DATA SCIENCE FOR MANAGERS

SECOND EDITION

HOW TO USE DATA (BIG AND SMALL) TO SOLVE BUSINESS CHALLENGES

RICHARD BOIRE

Data Science for Managers

Richard Boire

Data Science for Managers

How to Use Data (Big and Small) to Solve Business Challenges

Second Edition 2026

palgrave
macmillan

Richard Boire
Boire Analytics
Lakehurst, ON, Canada

ISBN 978-3-032-19704-7 ISBN 978-3-032-19705-4 (eBook)
https://doi.org/10.1007/978-3-032-19705-4

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Terms and subjects such as "Big Data", data science, predictive analytics, and artificial intelligence (AI) are commonly used in today's workplace, but are they really understood by the community at large? The field of analytics is growing at a significant pace, but it is not a new one. Much of the growth is due to our ability, thanks to computer technology, to capture, store, manage, and action vast amounts of data. Data science and analytics is being used in almost every industry and across every facet of our society. From understanding consumers and what they are likely to purchase, to identifying potential terrorist activities, predicting the likelihood of someone being a good employee, and understanding the probability of where a baseball player is likely to hit the ball, data science and analytics is everywhere—whether you notice it our not.

In today's world of AI, the notion of dealing with large volumes and varieties of data at rapid speed is a growing requirement as successful solutions require vast amounts of data. What to do with this data, however, is still a challenge for most organizations today. It takes experience and knowledge to weave through the "Big Data" maze to convert this raw data into powerful and meaningful data science solutions which, in many cases, will encompass the use of AI. In a time when technological capital has outpaced human capital, this book will help accelerate your learning curve.

In Richard Boire's book, *Data Science for Managers-How to Use Data (Big and Small) to Solve Business Challenges*, you will learn from a leader in the field of analytics. He has been a practitioner in the field of data science since well before such terms as "Big Data" and "artificial intelligence" became commonplace. This book will allow you to benefit from his 35-plus years of experience and take advantage of the important and vital lessons he has learned along the way.

Richard's practical and pragmatic approach to solving business problems is on display throughout the book. Leveraging real business examples, you will learn: how to evaluate or define the problem to be solved; how to leverage internal and/or external data to solve the problem; how to build, create, and

select the best tool or solution; and, finally, how to properly implement and measure the impact of the solution.

You will also benefit from his experience as an educator and a mentor to many. You will gain insights into what makes analytical initiatives successful. Effective data scientists can't work on an island. Successful data science initiatives require teamwork and interaction with other key disciplines, such as data engineers or data technicians/IT, and the end business user, often referred to as the "subject matter expert (SME)". And those most successful solutions require a combination of both art and science. This book is an opportunity for you to share in this learning and is a must-read for anyone new to, or experienced in, the area of data science and analytics.

Preface

The first version of this book on data science was written and published ten years ago with the title *Data Mining for Managers*, when the word data mining was more commonly accepted by practitioners. Of course, the discipline has evolved since then, and the more commonly accepted name today is now data science. Yet the principles, methodologies, and approaches remain the same. There have, however, been tremendous developments in the use of unstructured data for Natural Language Processing (NLP) and text analytics, as well as the use of artificial intelligence (AI) in the development of analytics solutions.

Besides my career as a data scientist, my role as a part-time professor allowed me to utilize the original version of this book as a reference source for a number of my teaching courses. Given all the developments in the past decade, however, I felt a compelling need to update the original version. This revised version of the book still provides the necessary foundation of knowledge in data science, which, as stated above, really has not changed over time. Yet at the same time it is my hope that it provides a perspective on how these newer developments in AI and unstructured data can be applied within the discipline of data science.

Lakehurst, ON, Canada — Richard Boire

Acknowledgments

In a revised edition of this book, there are, of course, those individuals, as mentioned in the first version of this book, who are once again mentioned in this edition. But with this new edition, I would also like to thank the colleges in which I have taught, George Brown College and Seneca College, for their support as I explored such topics as artificial intelligence, text analytics, and image recognition.

Environics Analytics, and in particular their leader, Jan Kestle, also provided more in-depth and detailed learning on the use of geo-demographics in providing better data analytics solutions.

A practitioner-based book like this can only be written on the basis of its author's real-life experiences. In this journey, there are indeed organizations and individuals I need to thank since without them the book would not have been written.

Reader's Digest and American Express were key organizations in my early career, providing the foundation and culture to build my learning in data mining. On a personal level, I would like to thank David Young, vice president of Direct Marketing at American Express Canada and a key mentor, for demonstrating how to socialize the benefits of data mining to key business stakeholders.

As the data mining discipline is still somewhat new and emerging, the knowledge base continues to evolve. It is my belief that a book that could leverage my learning and experiences as a practitioner would add to the knowledge base of this discipline. Of course, the most important organization in providing these experiences has been the Boire Filler Group. For 15 years, our team has been able to provide the benefits of data mining to organizations in all industry sectors. We owe our success to the valued and talented people, as well as the leadership, within the Boire Filler Group. No thanks would be complete without recognizing the leadership of this organization under my partner, Larry Filler. He has been instrumental in building our business but, more importantly, in allowing our organization to provide work in a variety of different sectors, which is the basis of most of the content in this book.

On a more personal note, I would like to thank my family, and specifically my wife, Clare. I would like to dedicate this book to her—as her undying patience and support have been the real catalysts for my success and enthusiasm in this discipline.

Competing Interests The author has no competing interests to declare that are relevant to the content of this manuscript.

Contents

List of Figures

CHAPTER 1

Introduction

The challenge in writing a book on data science is to differentiate its content from the plethora of other books, articles, seminars, etc. that have been written on this topic. The extreme popularity of data science is, of course, due to the recognition that most businesses now understand the value of information and use it to make more profitable decisions. In most organizations today, data science either is or will be a key business process. Since data science is a relatively new business notion, knowledge and understanding in this area is a critical need for its successful implementation within any business. Certainly, the consulting arena has grown in this area as more people profess to be experts in this area. With the increasing popularity and acceptance of artificial intelligence (AI) as an everyday facet within our lives, the heart of its success is data. Yet, this can only be successful through the discipline of data science which not only looks at mathematical techniques and algorithms but also explores the data that is fed into these algorithms.

But companies should first recognize that this is a new field, meaning that it is difficult to find practitioners with a vast array of experience within this area. Like with any other discipline, the key to becoming knowledgeable and acquiring expertise is to put ideas into practice and observe their results. Although much of this practical knowledge has been acquired within the direct marketing arena, the approach and processes regarding this discipline of data science are similar when applied to non-direct marketing areas. The internet, and its resulting explosion of digital data, have made data science a more mainstream topic within our world today as decision-making becomes more data driven while business intuition becomes a relic of the past.

The real key to attaining data science excellence is how one deals with the data. This is the fundamental essence of data science and one which the reader will realize through the course of reading this book.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_1

The Digital Impact

The seasoned and experienced practitioners will also tell you that the largest challenge facing them today is how to optimize the use of data science given the explosion of new software and technology that is currently available. The other significant challenge facing practitioners is the use of data science within the web environment. With both the web and technology, data science analyses can be conducted from the moment that a campaign is launched. Since information is readily available from the launch of a campaign, analyses can be conducted almost instantaneously. Access to this type of information has already created the demand for tools and technology which will facilitate both the volume as well as the speed of analyses. In the past, analyses used to take six to eight weeks to derive any learning. With the web and digital data, we can derive learning from day one. This kind of capability will also allow us to establish elaborate testing environments in which learning can be optimized.

Any discussion of the web is incomplete without acknowledging how processes have changed through Big Data technologies, and, of course, how we are now able to use AI in a more effective manner. This revised edition will explore these concepts in much greater detail.

Demand for Services

As the field of data science is still in its infancy, it is very interesting to observe the impact of this sector within business. Its overwhelming reliance on technology implies that the industry is quickly evolving in terms of both available tools and software. It will also place tremendous demand on those individuals who are true practitioners. The demand for those data science practitioners is now evident in any of the job listing websites. Within academia, new disciplines in data science, data analytics, and AI are being created both at the bachelors and graduate level. Yet, the challenge in these disciplines is that change is occurring at lightning speed due to a constantly improving technological environment. How do we effectively develop the data science practitioner of tomorrow? This is not an easy question to answer because, although there are concepts in this discipline which are unchanging, there are also other concepts dealing with the applications of mathematics and technologies in handling data that seem to be changing all the time. But the focus in the development of any data scientist should be the hybrid approach in which the practitioner can easily complement their technical and mathematical skills along with their business understanding and acumen. More about this concept will be discussed in later chapters of the book.

Art and Science

Although the technology is becoming more sophisticated and complex in terms of providing additional targeting capabilities, no one should doubt that there will always be an element of art in building data science solutions. Pure

understanding of the data without any cogent business understanding of the results will generate conclusions and recommendations which can be very misleading.

For instance, one retailer when conducting a data science analysis discovered that the sale of beer and diapers were very highly correlated. Does that really mean then that customers who buy beer are most likely to purchase diapers? Upon further investigation, it was found that these results were caused by the fact that beer and diapers were placed adjacent to each other within the store. As it turned out, young fathers were shopping for diapers late at night for their wives and, hence, decided to pick up some "suds" along with diapers. The important lesson here is that further investigation was warranted to truly uncover the "insight" from just the raw numbers. In this particular case, the data science analysis was skewed simply by product placement within the store rather than consumer intentions. As data scientists, we can determine the correlation between variables or characteristics but not causation. This is eloquently discussed in **Eric Siegel's book *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*.** In the book, he provides an example that, according to correlation analysis, people who eat ice cream are more likely to be eaten by sharks. Impact is there but not causation as one could look at a sequence of events which more fully explains the relationship. People tend to eat more ice cream on hotter days. They also tend to go swimming on hotter days. And they will tend to swim more in the ocean, where there is that likelihood of being bitten by a shark.

The art component also allows the analyst or data scientist to better interpret results. With a solid understanding of the business, the data scientist can better understand why these results are occurring. At the very least, he or she can better investigate certain facets of the business which may provide the underlying rationale for why these results are occurring. Although data science is number-intensive, it is this complementation of business domain knowledge, in conjunction with the numbers, that allows the practitioner to deliver more optimal results.

Data science is not solely about technology; it is also concerned with the use of technology to arrive at business solutions which ultimately optimize customer return on investment (ROI). This requires organizations to invest in intellectual capital as well as technology. In fact, successful organizations will prioritize their investment on the intellectual side rather than on the technological front. Utilizing this complementation of technology and intellectual capital will truly enable companies to optimize customer ROI. The key to the successful use of data science is to always produce solutions which can truly optimize ROI at its most granular level which in most cases is at the customer level.

Data science continues to evolve both as a science and an industry discipline, yet one must recognize that there is an element of art if one is truly to be successful as a practitioner of data mining. At the same time, the value of data science is now widely recognized as contributing huge benefits to our society.

For example, within the business sector, the ability to target the right product to the right person at the right time results in lower marketing costs, but often yields incremental revenue. This type of economics often translates into ROI increases of 100% or more. The credit risk/fraud area, which is often considered the birthplace of data science, as we will see in later chapters, attempts to identify individuals who are likely to default on payments or in the case of fraud, identify individuals who exhibit out-of-pattern spending. Improvements of 1% in the areas of credit risk and fraud due to data science can translate to millions of dollars, which directly impact the bottom line.

Industry Perspective

Health

In hospitals and the health sector, we see health professionals analyzing a vast array of data and information to identify patient problems or diseases. By having access to the patient's history as well as perhaps hundreds of other patients' histories, the health sector is better equipped to identify the given challenge or problem, even in relatively unique situations. This access to tremendous amounts of data can often lead to a unique treatment for a given patient, which is, of course, dependent on the particular problem and the history of the patient. Utilizing this vast trove of patient data, healthcare institutions can predict the costs that will be incurred in managing the patient back to health. If an institution can predict that I will be in the hospital for two weeks compared to the prediction of two days for my wife, then the hospital could prescribe a set of activities with the desired goal of reducing my hospital stay to a week or less.

The value of data science in today's medical environment is that we can now analyze and marry vast amounts of data pertaining to patient histories as well as problems with the intention of providing the optimum course of action for a given patient and problem.

Government and Law Enforcement

In government, we are seeing the use of data science as a law enforcement tool. If we think about data science as a knowledge discovery tool, we can then understand that the key to knowledge discovery is the ability to detect unique patterns which reside within the data. In law enforcement, we all understand the notion of "detective work" as detectives personally deal with their own repositories of data. These repositories of data consist of notes in a handbook, previous experiences, as well as hunches. As one can see, the acquisition of this type of knowledge is manually intensive. From this knowledge, the detective attempts to identify the unique pattern which will result in a potential arrest. With data science, a higher level of automation can be utilized to the extent that vast amounts of data and information from a variety of detectives are

collected from one crime scene as well as perhaps many other similar crime scenes. The data is then compiled into one analytical file or database and then statistically or mathematically analyzed to uncover unique patterns of behaviour or occurrences, which could help to solve the crime at hand.

Besides solving a specific crime, the use of this data may also help to forecast which areas within a city are most prone to specific crimes. With this type of information, the police can then better allocate their resources. This smarter allocation of resources occurs because the city can send both the right quantity as well as the right type of officers, based on the type and volume of crime that occurs for a given area. For example, this is exactly what New York City did under Rudolph Giuliani in their attempts to reduce the horrific crime rates and, in particular, the homicide rates that were occurring in the late 1970s and early 1980s. New York's homicide rate is now down to less than a third of what it was in the late 1970s, and the use of data science technology is recognized as one of the key reasons for this significant decrease.

The notion of using data science as a law enforcement tool has certainly gained even more prominence in a post-9/11 environment. As one of its many its responses to the terrorist attacks of September 11th, the U.S. Government introduced the development of a department to oversee the compilation of private individual data from various government organizations and its use in combatting terrorism. This massive database could contain information pertaining to areas such as health, insurance, banking, etc. Essentially, the government could have a dossier on everyone's activity and behaviour over the course of that person's life. Besides the obvious privacy concerns, one of the arguments against such an exercise is that the development of tools and solutions for preventing terrorist activity requires that we have enough observations. In the case of 9/11 or the prevention of future 9/11s, the argument is that we would not have enough data to build profiles of information around Al-Qaeda terrorists, since there would have been no prior history to 9/11. Even with the 9/11 event, we only have 19 data points (19 terrorists), with which to contrast their information from the remaining 300 million non-terrorists in North America.

Because of the sparsity of data regarding individuals who partake in terrorist activities, the other approach is to adopt the data science tactics used in identifying fraudulent behaviour. For example, can we identify out-of-pattern behaviours, such as phone calls between certain groups of individuals? Indeed, there is the problem of false positives, where we identify someone as a terrorist when indeed they are not. But this error is far less egregious than the false negative error of not identifying someone who is a terrorist. However, the ability to identify someone as high risk can provide a trigger for associations/institutions to introduce a series of actions and steps on these identified high-risk individuals. The key here would be the establishment of those series and actions that would mitigate any negative fallout from those false negatives.

Various debates and discussions have also centered around the type of data being used to identify potential terrorists. A fallout of these discussions has

been the whole notion of "racial profiling" and the obvious sociological implications that accompany this practice. Yet, issues as sensitive as racial profiling will need to be addressed by examining how to balance the need for public security versus the need for protection against law enforcement discrimination. A few key stakeholders will need to be involved in these thought-provoking debates and discussions along with our government institutions. Ultimately, I hope that this process will result in a policy that clearly lays out the guidelines regarding the use of racial profiling within the context of law enforcement.

Using Data to Determine Optimum Site Locations

Besides the use of customer data and the utilization of loyalty programs to target customers, which will be discussed later in the book, another use of data is to determine the best site location for a given company. Here, data is used to identify the top-performing site locations. Then the data is used to determine the geo-demographic characteristics (i.e. age, income, ethnicity, gender, etc.) which best describe these top stores, or generate a profile of the best locations. This information is then used to identify those areas where there are currently no stores, but which have a high-profile score.

Real-World Experience of Data Science

When I began my career as a direct marketer in 1983, directly after graduating with my MBA, I was fortunate enough to apply my academic knowledge particularly within the area of statistics. As with most young graduates, I was confident that I would make a difference within my new company. Yet, although I contributed to the organization in fulfilling the duties expected of my position, the organization contributed far more to me in providing a basis upon which to build my career. This basis was grounded in the notion of applying statistics but within the practical arena of business. This meant that the esoteric and arcane principles of statistics would be followed but not adhered to in the strictest sense. For example, in most cases, the traditional assumptions and rules of normality regarding the statistical analysis of sample groups within the analyzed group are not followed because most of the analytical situations using statistics within the business environment deal with non-normal populations. Yet business continues to apply these techniques much to the pure mathematician's horror because they work and produce acceptable results.

This initial learning about "bending the rules of statistics" was perhaps my biggest indoctrination into the application of statistics within the world of direct marketing, where direct marketers were the early pioneers of applied data science. In the world of academia, we were often acclimatized to regression results that produced R^2 of.7 or more, which is a statistical measure used to describe the performance of a given solution. More about this topic will be discussed in later chapters. Within the world of direct marketing, it was not uncommon to observe R^2 results of .05 or less depending on the data. Yet

these statistical results, which would be totally unacceptable within the academic statistical arena, were commonly applied in the business world. The ultimate reason for this divergence in opinion between academia and business is that both sides viewed the success of a given solution in very different fashions. We will discuss this in more detail in the later chapters.

As with any book involving a topic of great significance, it is important to differentiate how this book will be different from the others. If readers are looking more for the mathematical and technological nuances and insights regarding data science, then this book is not for you. However, if the reader desires a more practical business perspective of data mining, then this book is very appropriate. Using this type of perspective, I have decided to focus on data science from a practitioner's viewpoint. The advantage and benefit to readers is the 35+ years of applying several tools and techniques to almost a limitless number of business situations. Hopefully, my insights will shed light on what works and what doesn't under certain conditions. As with any book on data science, there will be discussion about some of the technology and mathematics that is employed, but always with a view of how it meaningfully impacts the business. Yet in the end, it is not my intention to convert all readers to become data science specialists. Rather, it is to impart insights and knowledge on how data science can shift the key levers of your business and, more importantly, understanding how this drives ROI. For those of you with lots of data science experience, this book can be thought of as another "point of view" which can be referenced by the experienced practitioner. Another advantage to this book is that it provides insights into data science but within a Canadian context. I recognize that there may be many books on data science, with most of them being written in the U.S. By adding a Canadian dimension to the discussion, it is my hope that this will enhance the learning and insights of the reader. After reading this book, if you the reader can use this as a reference regarding how specific data science tactics impact a particular business problem, then I will have achieved my goal.

Enjoy.

CHAPTER 2

Growth of Data Science and the Data Scientist

In writing a book on data science, it is always important to understand its history. What were the challenges and developments of the past that catapulted data science to the forefront of present-day business? To truly obtain a sense of its history, it is always useful to speak to its earliest practitioners, which are the direct marketers of the large major catalogue or publishing house companies.

As a recent MBA graduate back in 1982, I was fortunate to be hired by one of these pioneering direct marketing firms—Reader's Digest. At that time, I was hired to work in their list selection department as a regression analyst. Although we did develop predictive regression models for direct mail campaigns, we were responsible for all analyses or segmentations which dealt with individual customer-level data. Being the customer knowledge gurus of Reader's Digest and seeing the huge return of investment (ROI) payback of this type of business culture, I thought that this type of business process was normal for most organizations. Conversations with my MBA grad colleagues who were working at various industry-leading institutions in the early 1980s revealed that none of this type of work was being done. We had no fancy titles such as business intelligence specialist, knowledge discovery analyst, data scientist, or Customer Relationship Management (CRM) business analyst because these were the real "pioneering" days of data science and predictive analytics.

In understanding the tremendous business potential of data science in improving overall ROI and the fact that the Digest, along with only a handful of other organizations, were doing this kind of work, it became apparent to me that tremendous entrepreneurial opportunities existed within the area of data science. However, the mitigating barrier to becoming an entrepreneur in this field was technology. Tremendous costs in computing were required in order to capitalize on some of our statistical techniques, which were employed for various direct marketing campaigns. In 1982, there were virtually no PCs, so all our work was done within a mainframe environment. Yet, the analytical

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_2

environment was very robust and efficient within this organization. At Reader's Digest, we had extensive campaign history for each customer. We knew when they were promoted, how often they were promoted, the types of promotion, and how often they were promoted since their last purchase. With this promotion behavior and purchase behavior, we were able to develop robust models quickly. We had multiple models for each product line based on their specific promotion history. For instance, we had specific One Shot Books (OSB) model for 0–3, 4–6, and 6+ OSB offers since last-order segments. Furthermore, for each product promotion history segment, we had multiple models built for specific times of the year in order to take account the impact of seasonality on campaign performance.

In many cases, our segmentation strategy and schema were developed from the past results and learning of prior campaigns. By using our business intuition and judgment on these past results, we were able to make sound business decisions regarding any new business segments. The use of statistics was employed if it was clear that incremental business results would be achieved. Reader's Digest was always proactive in testing new technologies and approaches in order to be at the forefront within the direct marketing world. This extended to the world of being able to better target customers rather than being able to simply merge purge lists or personalize names within a direct mail piece.

The use of statistics was judiciously applied in the sense that our methods for evaluating different statistical techniques within a live business initiative were not always perceived from the statistician's purest viewpoint. For instance, in the academic world, when evaluating multiple regression techniques, it was not unusual to witness R^2 about 70%–80%, where 100% represents perfection. Essentially in layman's terms, this benchmark attempts to measure the amount of explained variation produced by the model relative to the total variation of the data. (Note: the concept of R^2 will be explained in more detail in a later chapter.). Yet, in a direct marketing campaign, it was quite normal to observe R^2 ranging only from 1%–4%. Even with these depressed R^2 results, many of these campaigns were largely successful because of their predictive models. In the real business world, large random variation is the norm rather than the exception. By trying to explain just a small component of this variation, the use of predictive models yielded significant business benefits to Reader's Digest. The discipline of measuring techniques and their impact on live business data is critical to any evaluation process. As a result of this discipline, Reader's Digest was able to enact a business culture in which statistical learning was always complemented with business learning.

As with any new business process, the right culture needs to exist if data science is going to produce successful results. Positive mindsets amongst employees, and the willingness to embrace change and new learning, are critical components of this kind of culture. Hiring the best people and purchasing the best technology will produce less than satisfactory results if the organization itself does not have the right fit for these types of practices. Having the right "fit" in a sense speaks to the corporate culture of the organization. We will

discuss the type of corporate culture which is necessary to make data science an ongoing business process within the organization. Presuming that an organization is starting at ground zero, how does an organization evolve to the point that it becomes a leader in the practices of data science? As with many things in business and in life, it is about incremental steps. In data science, incrementality provides a continuous stream of knowledge that ultimately reinforces the discipline of data science within the organization.

For marketers, it is in their nature to deal with acronyms and buzzwords when describing the latest ad or campaign. It is certainly no different when we describe a process or culture which ultimately changes the way we both think and execute within the work environment. This is certainly the case with data science. A multitude of books and articles have been written on this topic, and there is certainly no shortage of people who claim to be experts in this field. Given the number of seminars, white papers, books and courses, it would appear that this discipline is experiencing exponential growth which is reinforced by the explosive demand for data scientists. Yet, the reality is that data science is still in its infancy. Why is that?

As with other changes which are occurring in our society today, technological advancements have created an environment in which more players can play the game. In the late 1970s and early 1980s, only the large direct marketing businesses, such as Reader's Digest, practiced data science. They were able to conduct these practices by investing millions of dollars into the development of appropriate systems and databases required to process the data. Certainly, the investment in this technology was justifiable since direct marketing was its core business.

Direct marketing often gets a bad rap as being simply "junk mail". But the principles of successful direct marketing are what data science is all about. For example, the ability to uncover knowledge about customers and apply this learning to future business initiatives is core to the success of any direct marketing or CRM initiative. Reader's Digest always knew who to promote, what to promote, and when to promote. However, for most other organizations, this was not the case, as direct marketing represented only a minor portion of the typical marketing budget. The business paradigm for most marketers during the late 1970s and early 1980s was an optimization of revenues which was supported by the prevailing business culture for that time.

Yet the technological advancements in the last 25 years, particularly with Big Data and artificial intelligence (AI), have changed how society functions as well as its expectations. With technology, we are expected to do more with less resources.

Consequently, consumers are now bombarded with all kinds of communications and offers in the hope that they will respond appropriately to each company's message. With the Web becoming an ever-increasing marketing channel, we now have the ultimate medium whereby the collection and actioning of individual information can occur dynamically. This means that decisions regarding how we communicate to customers in this medium can change

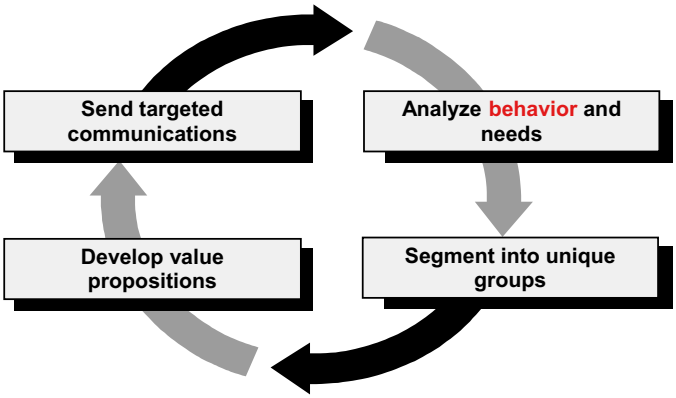


Fig. 2.1 The lifecycle of analytics and decision-making

instantaneously. In the old direct marketing environment, we had to wait for several weeks after the launch of a campaign before we received any information from consumers. The dynamic and interactive nature of the Web requires that data science become a more firmly entrenched discipline.

The increased access to information and, more importantly, the ability to use it is a significant competitive advantage in today's business environment. Organizations operating under this type of knowledge environment can view all their business initiatives from a ROI standpoint. At the same time, these organizations can gain tremendous insights into what worked and what didn't. This dynamic learning process provides an environment whereby the business is always striving for improvement. This is the key to successful data science. The cycle of campaign execution and learning provides the ongoing feedback to constantly improve future marketing efforts (Fig. 2.1).

The best way to illustrate this is through a real-world example. Although example dated (prior to 2000), it is an excellent demonstration of how data science evolves over time to create both increased knowledge as well as its resultant increased analytical demands.

Case Study: American Express Canada

In the 1980s, American Express was an organization striving to increase its membership base. In fact, its share price was directly impacted by how many new cards were acquired, referred to as Cards-in-Force (CIF). Towards the end of the 1980s, CIF targets were being achieved but at a rate where the cost per new card doubled. These cost inefficiencies were unacceptable to the organization. To resolve this dilemma, they began to purchase names from lists which seemed to fit the profile of an American Express cardmember. Lists such as Business Week, Forbes, Fortune, etc. produced very good acquisition results. However, the number of names on these lists were simply too small to provide

the required prospect universe which was necessary to generate Amex's aggressive CIF targets. Ultimately, Amex needed a very large prospect universe while using tools which allowed them to generate new cards in a cost-efficient manner. The solution was twofold.

First, they built a prospect history database which compiled information at the individual prospect level. Examples of some key information included the promotion history and gender of the prospect. With this database, Amex began to bucket its prospects into two key segments based on promotion history: new names and previous names.

The second solution was to then build predictive models which could be applied to both universes to optimize response rates. Using overlay aggregate-level data from Statistics Canada, gross response models were built and provided a 50% lift in acquisition performance. However, learning revealed that approval rates had deteriorated. They had indeed discovered that the best responders were those most likely seeking credit. The learning from this campaign indicated that Amex needed a targeting tool which optimized net response rate (gross response and approval) and not just gross response. A net response model was developed and, once again, they achieved a 50% lift in acquisition performance. However, they achieved much better cost efficiencies within the new applications area since these applicants were more creditworthy. This ultimately resulted in lower credit losses for this group of customers within their first 12 months as an Amex customer. Yet, the learning also revealed that, though, they may indeed be optimizing net response rate, these new cardmembers may not be profitable. In other words, they may not be using the card or, worse, simply cancelling it. As a result, Amex decided that the goal of acquiring new cards should be based both on their net responsiveness as well as their overall profitability within their first year.

Solutions predicting response, attrition (cancel), and credit risk can, in essence, be combined to provide an estimate or predictive value for an approximation of customer profitability. The word approximation is used as the real notion of profitability implies that all fixed and variable costs and revenues are factored into the definition. Adopting this approach of calculating profitability in its truest sense consumes an inordinate amount of time and, in fact, is counterproductive to the goal of what we are trying to accomplish. The goal in this exercise was not to arrive at an exact definition of customer profitability but rather to produce a more relative measure of profitability. In other words, the goal was to arrive at a profitability metric that could differentiate customers based on a relative measure but not to provide an exact or absolute measure. Even so, this required buy-in and consensus well beyond just the analytics department. The finance, marketing, operations, and systems departments were all involved as key stakeholders to ensure that this so-called "profit" metric be recognized as a key measure when engaging with customers. At the end of this process, consensus was obtained amongst all the key stakeholders in arriving at this quasi-like "profit" metric. Recognizing that the profit measure was relative rather than absolute, this measure was used to assign customers to

different segments with the very high group being the most profitable and the very low group being the least profitable.

Now that agreement was obtained on how to calculate these metrics, the next part of the exercise was for the analyst to create this measure at the individual customer level using existing analytical tools. In this case, SAS was used as the analytical tool and reports were built to see what this measure looked like when compared to other key business indicators (KBIs). An example of one type of report is listed below (Fig. 2.2).

Looking at the above reports, the profit metric does a really good job in being able to assess retention performance, customer spending, and credit risk based on the rank ordering of these KBMs against profit. The profit metric also performs adequately in assessing the average time to make payment, and the average number of products, although it is not as pronounced as the other KBIs. This kind of report provides valuable information to all the key stakeholders in determining whether or not the profit metric is a meaningful measure of profitability.

With the profit metric being defined and agreed upon by all the key stakeholders, models could then be built to predict profitability. At this point, we then had models to both predict net response as well as profitability. Using these above models, prospects could be evaluated based on

$$\text{Predicted ROI} = \frac{\left(\text{Predicted Profitability}\right)}{\left(\text{Predicted Net Response}\right)}$$

% of Customers Ranked by Customer Profitability	Index of Avg. Profitability per Decile	Index of Average Retention Rate per Decile	Index of Average Monthly Customer Spend per Decile
0-10%	5	4.5	4.2
10-20%	4.2	3.9	3.7
20-30%	3.2	2.8	2.6
30-40%	2.7	2.6	2.2
...			
90-100%	0.3	0.3	0.5
% of Customers Ranked by Customer Profitability	**Index of Average Credit Risk per Decile**	**Index of Average Time to Make Payment**	**Index of Average # of other Products**
0-10%	4.4	3	2.8
10-20%	4	2.6	2.4
20-30%	2.9	1.9	2
30-40%	2.7	1.6	1.5
...			
90-100%	0.5	0.75	0.7

Fig. 2.2 Example of Key Business Measure (KBM) report and customer profitability

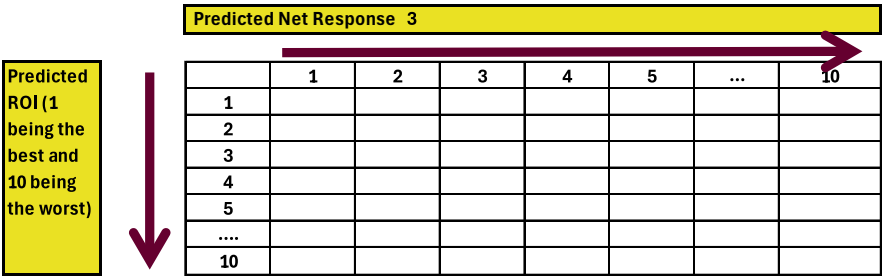


Fig. 2.3 Example of 10 × 10 gains chart matrix

In this exercise, which encompassed several years, the work itself was able to produce the following outcome and decision report (Fig. 2.3).

This allowed the Amex marketing team to select prospects on ROI as well as net response, as one of their key performance measures was CIF. In effect, the use of this decision tool allowed marketers to demonstrate the impact of optimizing ROI versus lost cards opportunity.

This case study demonstrates how an organization evolved their overall acquisition strategy based on data science. The decision-making process in each stage was based on results and provided input on what the organization needed to do in order to achieve the next level of improvement.

The final outcome of this multi-year project, which evolved over time, was the ability to significantly increase our decision-making process in selecting the best prospects that would become profitable Amex cardmembers.

Key Takeaways

- Data Science is not a new discipline as direct marketers were the early pioneers of data science and machine learning.
- These early adopters demonstrated the importance of culture in embracing data science as a core corporate discipline.
- These early pioneers recognizing the tremendous benefits achieved through data science invested huge amounts into technology and computer systems.
- Data science is a cyclical process such that we have action, execution, and learning where learning feeds back to action to improve the overall process
- One example of an early adopter was American Express, who demonstrated the tremendous business benefits of data science as an ongoing discipline.

CHAPTER 3

Data Science in the New Economy

It is important to understand that the same learning feedback process used to generate strategy improvements can also be used in a more granular fashion by data scientists in identifying specific tasks and tactics which can be improved for future business efforts. Besides the notion of ongoing learning, successful data science practices include the application of this learning for future activities. The adage of "Analysis for Paralysis" is a trap that businesses can easily fall into whereby nothing is ever actioned. Here, businesses in an attempt to be competitive through data analysis fail to understand how any of its resultant learning might be actioned. The objective of any marketing environment is ongoing learning, which can only be supported through continued execution of prior learning. As we saw earlier in Fig. 2.1, this principle of an ongoing feedback loop of learning represents the underlying philosophy behind any data mining exercise.

However, like with any business process, the key is having the right people in the right positions. This certainly applies to the field of data science. Any investment within this sector should focus on people first followed by technology. Yet, technology continues to increase at an exponential rate, especially with artificial intelligence (AI) becoming an everyday common business practice. The need for the data science hybrid is even more paramount as companies seek individuals that can use both technology and their problem-solving skills when undertaking a project. Technology is no longer a barrier or limitation in providing a solution, as we can now solve more business problems. In many ways, the new paradigm of the data scientist is that of a generalist, unlike the early days when they were seen more as specialists.

Definitions abound about what data science is and what it isn't. However, with the many experts and pundits offering their opinions and views on data science, not much discussion has ensued concerning the roles and responsibilities of people within this field. We all know that the field of data science has

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_3

some impact on virtually all areas within any business. With increasing data accessibility, intuition is a relic of the past as organizations become more data-driven.

But what about the role of the data scientist? This is still considered a relatively new position for many organizations, given that this role was unknown prior to 1990. Yet, many, if not all of the tasks now outlined within the discipline of data science were being executed by individuals who were working at traditional large direct mail companies, such as American Express and Reader's Digest. Instead of being called data scientists, they were typically referred to as modeling/statistical analysts.

The data scientist ultimately derives his success by serving as the conduit or hybrid between the technology area and the business functional area. Traditionally, the key functional area was marketing. In today's environment, one could argue that all departments within an organization look to data science as the key competitive advantage in becoming more data-driven.

Although the ultimate career goal of data science success is to become that hybrid, the initial skillsets of data scientists, when first commencing their careers, tend to focus more on the technical expertise. With technical expertise as a foundation, organizations often attempt to identify those individuals who also have the acumen and adaptability to grow into a more generalist type of role. This is not an easy challenge, as the generalist and technician wear two different hats. Consequently, successful organizations will ultimately hire those persons with the right balance of both technical and general skills.

These seemingly contradictory skillsets are best understood by describing some of the roles and responsibilities of the data scientist. Within a given business situation, a challenge is presented to a group of key stakeholders. The data scientist needs to understand, first and foremost, where data science can have an impact. For example, one company may have retention rates of 10% for Product X and 90% retention rates for Products Y and Z. It would appear that the company has a much larger issue than the data science issue of identifying those customers who are most at risk of cancelling Product X. In fact, the priority would probably be to better understand the entire product offering, rather than focusing on data science efforts for better targeting of high defection risk customers.

Conversely, Company DEF has experienced tremendous growth in acquisition over the last 5 years, yet costs per new customer have significantly grown by 20% in the last 2 years. In this case, the data science issue of reducing costs per new customer through better targeting tools would apply.

You can see from these simple examples above; solid business generalist skills are required to identify both the opportunity and challenge which data science can impact. With the challenge identified, the data scientist now needs to wear a technical hat by identifying the appropriate data sources, technologies, and technical tools which are necessary for a given solution. Once the solution is developed, the application of that solution requires that the data scientist again wear two hats simultaneously. For instance, the business generalist hat is worn

during the design of an overall testing and tracking matrix to ensure that both learning as well as business performance objectives are being optimized. Yet, the technical hat is worn in applying this data to build this matrix.

This new emerging role of the data scientist is gaining increasing popularity as being a critical function in today's Big Data world. The demand for data scientists to think in depth about a business or organization is a fundamental prerequisite in developing solutions. The technical skills may be in great abundance, but how do data scientists create the right journey or path in solving these problems unless they can amply demonstrate the soft non-technical skills of understanding the specific entity or business? In other words, the ability to increase one's domain knowledge about a business or entity enables the data scientist to build a better solution.

Technical skills alone are inadequate as it is these hybrid-type skills that are the key elements in being successful as a data scientist. Listed below is a sample of some of the questions that the data scientist would ask in order to acquire more domain knowledge of a given project or exercise

1. What are the overall objectives?
2. What is the desired project solution, and is it tactical or strategic?
3. What are the limitations of the data environment?
4. How we will construct the appropriate data environment?
5. What are the expectations?
6. What are the tools and technologies that we will deploy?
7. How will we apply the solution?
8. How will we test and track our results?
9. What is the revenue model?
10. How would you define success in the completion of this project?
11. Are expectations equivalent amongst all the project stakeholders?

We are all, to some extent, experiencing the drastic transformations that are underway within our current economy. The advances in Big Data and AI are causing huge disruptions in the economy, with the result being that significant job losses are occurring in many sectors. In fact, certain sectors such as manufacturing, have experienced significant job losses that are gone forever. This situation is also now being experienced in many of the white-collar professions through global outsourcing and now AI. Of course, economic transformation is nothing new as economic historians can recite numerous time periods when the economy underwent major changes. Besides this economic transformation and the requirement for corporations to become more knowledge-based, we are also witnessing social transformations. In particular, tremendous emphasis is now focused on the environment and how we, as a society, can be responsible stewards of its resources. The more traditional disciplines that were primary users of data science, such as direct marketing, would argue that their targeting efforts enable businesses to consume less resources by sending out fewer mail pieces. With AI, though, we are now hearing about its great energy consumption needs and its obvious climate impact.

Alongside the need to become more effective at using environmental resources, the new economy has also resulted in fewer corporate resources being available to do a given task. Once again, the notion of "doing more with less" is being reinforced, with data science being the key discipline, despite the obvious ethical concerns.

These economic and societal changes are resulting in demands for skills that focus on the ability to explore and analyze information into meaningful business intelligence. As long as there are problems to solve and access to data continues to increase, data science jobs will represent areas of growth in the new economy. But, as will be discussed in later chapters, the job requirements and responsibilities will evolve more towards the generalist than the specialist.

Measurement Is a Critical Output from Data Science and Analytics

Even in economic downturns, discussions with data science counterparts have indicated virtually no decrease in their work activities; indeed, in many cases, there has been an actual increase in their workload. None of this should be all that surprising if we understand that one of the demands from the new knowledge-based economy is accountability. Marketers, who have been the pioneers in using data science, have always stressed the need for accountability and measurement, with the marketing mantra being "What gets measured gets done". Establishing a return on investment (ROI) on every marketing initiative is now a common discipline for most marketers, with data scientists representing vital components of the marketing team. The data scientist's ability to understand data, along with the marketing team's business needs, enables organizations to create ROI templates for a variety of marketing initiatives.

Data Science Solutions to Optimize ROI

In addition to providing a level of accountability through measurement, another demand of this new economy is the ability to create solutions that actually maximize ROI, rather than the focus being on revenue maximization, which was the traditional marketing objective through most of the twentieth century. The data science industry itself had traditionally been focused on the development of predictive models seeking to maximize revenue given a certain cost (e.g. response rate), which can ultimately be translated to ROI.

For many organizations, data science is still a daunting term. Although virtually all organizations recognize its importance as a key business discipline, many organizations lack the ability to truly assess its value from a quantitative perspective. How can organizations change this? Like most business roadblocks, the key is to identify quick wins that demonstrate significant positive benefit within a very short timeframe.

Customer Value as a Quick Win

Let's look at some examples of quick wins. The first and often most common quick win is for an organization to identify its best customers. Using the principle of the 80/20 rule, we attempt to see how company revenues are distributed amongst its customers. Using purchase revenue (which for most organizations is more than adequate in identifying customer value), we can create a measure of a customer's value by summing up each customer's purchase revenue over the last 12 months. Customers are then sorted into deciles (10 equal size groups of customers), with decile 1 being the highest-value group and decile 10 being the lowest-value group. Figure 3.1 is a schematic of what this might look like.

The production of this type of report (often referred-to as a "gains table" or decile report) reveals that 60% of the company's revenue is being delivered by 20% (deciles 1 and 2) of customers (not quite the 80/20 rule mentioned above). In this instance, customers in the top two deciles would represent our highest-value customers. Medium-value customers would be represented in deciles 3 and 4, where they account for 25% of all revenue. Finally, low-value customers are found in the remainder of the file (deciles 5–10) and account for only 15% of all revenue. This information is critical in enabling organizations to effectively deploy customer management strategies and tactics. The basic ability to prioritize marketing activities against different customer groups based on customer value can yield significant performance returns. For example, a greater proportion of focus and resources should be devoted to those higher-value customer segments which can potentially deliver the highest returns. The end objective here is alignment of company resources based on customer value.

Customer Decile	# of Customers	% of all revenue captured in decile	Segment Group
1	50,000	35%	High-Value Customers
2	50,000	25%	
3	50,000	15%	Medium-Value Customers
4	50,000	10%	
5	50,000	5%	Low-Value Customers
6	50,000	3%	
7	50,000	3%	
8	50,000	2%	
9	50,000	1%	
10	50,000	1%	

Fig. 3.1 Value segmentation report

Change as a Quick Win

Another quick win looks at the notion of changes in customer behavior. Marketers historically have been interested in being able to identify changes for a customer and, in particular, *lifecycle* changes. The key, of course, is to be able to identify these lifecycle changes based on what we can observe from customer behavior and demographics. Knowledge of a given customer who is a university student and is about to graduate would represent key information to a marketer in being able to design the appropriate programs that might appeal to this person. Suppose we observe that a person has just turned 65 and we notice that certain types of expenditure have significantly changed. For example, expenditures related to gas purchases have decreased while purchases related to travel have increased. You could certainly impute that this customer may have become a retiree. In other examples, we might see a married person significantly increase purchases on categories related to furniture. This might suggest that this person is buying a house. With this same married person, we might observe that purchase activity has now significantly increased in areas such as baby clothing, implying that this same person has now become a parent. In all of these change examples, data is required to make inferences like those above. Inferences can be made with any data, but the type of available data will dictate how confident we are in our inferences and how we manage the expectations of those using the insights. For example, a change in an investment portfolio whereby a customer aged 30 has gone from a riskier portfolio to a less risky portfolio might suggest that the customer might be getting married, starting a family, or simply just changing their investment strategy. In any case, the marketer would develop strategies and tactics that could leverage these insights. In all these examples, the use of analytics in obtaining insight from data is negligible as the data itself provides the insights.

Engagement as a Quick Win

Change and value both represent good quick win areas to focus on from a data science perspective; another area relates to customer engagement. Although engagement may be considered to be related to value, it is not always the case. A high-value customer at a bank may have a large loan and large mortgage yet make all payments through automatic withdrawal. If this is the way the customer interacts with the bank, the customer's engagement level may be very low. We want to consider other types of information in order to obtain a true reflection of customer engagement. With the growth of digital marketing, the extent of customer engagement becomes much broader as we can now look at customer-initiated activities such as email communications, website browsing activity, texting as well as in-bound telephone calls, and other means of customer-initiated communication with the company. This communication-initiated activity, complemented with purchase activity and response activity across all channels, can be integrated to create a more complete measure of customer engagement.

In today's new economy, data science and analytics is quickly becoming a common business discipline. Given that this discipline is still relatively new, most people tend to consider data science and analytics as simply the application of more advanced statistical techniques. It is this perception that has created a certain degree of complexity regarding the use of data science and analytics as a regular business discipline, when indeed simple solutions using the data in a pragmatic way often deliver great business value.

When assessing and prioritizing which projects to undertake, one should bucket projects into what are often referred to as the "low-hanging fruit" or those exercises which can yield great benefit. The "low-hanging fruit"-type projects represent exercises where almost nothing related to data science has been done in the past. Projects like this can achieve significant benefit within a very short timeframe, and often with minimal resources. The second scenario, of "high-hanging fruit", can also achieve significant benefits, but these types of exercises require more resources and commitment in order to achieve the desired solution. In most cases, practitioners are working in projects where something related to data science has already been done. In these cases, we are simply trying to achieve incremental benefit over and above an existing data science solution, as well as to maintain the benefit experienced when the solution was applied for the first time.

Success is primarily reliant on people. Nowhere is this more applicable than having the right team for a given project. Projects will be successful if we take a multidisciplinary approach whereby data scientists have expertise in areas such as mathematics/statistics, databases, and the domain knowledge of the business area.

The ability to both organize and manage this information is another critical element to success. The benefit of using high-powered mathematical solutions is only as good as the inputs (data and information) which are fed into these techniques, even with advanced techniques such as AI.

As with any data science project, practitioners must always be cognizant of its cyclical nature. This implies that the practitioner must be looking not only at maximizing results but also at how to obtain more learning once a given solution is applied within a specific business initiative.

No project will be successful without a "business champion". This person is not the "data science practitioner", but typically a key individual in the business who takes ownership of the project results. Having such a vested interest in ensuring that data science is successful, this individual in a way becomes an "evangelist" for these type of solutions throughout the organization. He or she becomes the ultimate champion of this type of solution.

Key Takeaways

- A need for a culture of ongoing learning through data science.
- This creates the need for a more hybrid type role in data science.
- Having business skills and technical skills provides more creativity in one's ability to solve business problems.

CHAPTER 4

Using Data Science for CRM Evaluation

As stated in earlier discussions, data science is not just about building tactical solutions towards specific issues such as the creation of machine learning/AI solutions. Data science can also be used to develop overall strategies such as the development of a broad segmentation strategy. We will talk more about segmentation later in this book. At an even broader level, data science can be used to determine whether the deployment of a Customer Relationship Management (CRM) program makes sense. It is the responsibility of the data scientists to build the case for support or no support of the given program.

Evaluating CRM in General

In this scenario, the use of a pilot project or proof of concept can provide the necessary information in a very quick timeframe and with minimal investment.

At this basic level, the goal is to determine whether a CRM strategy will yield a positive return on investment (ROI). Results from a small pilot test can be extrapolated and eventually rolled out as a CRM strategy. The following is one example of this process where we looked at the utility of retention programs:

Company ABC was unhappy with overall retention. It was looking at such options as the viability of a CRM retention program for their business. In the pilot project test, results showed that a retention program can increase retention rates by 2%. This then became the key metric in creating the spreadsheet below. Without a retention program, the retention rate was 85% while with a retention program, the retention rate increased to 87%. The other assumption of profit per customer of $300 was held constant both with and without the retention program (Fig. 4.1).

In building this spreadsheet, one might consider a short timeframe such as one year in which to measure the impact of a retention program. However, in

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_4

No retention program	Year 1	Year 2	Year 3	Year 4	Year 5
Number of Customers	10,000	8,500	7,225	6,141	5,220
Average Profit/ Customer/Year	$300	$300	$300	$300	$300
Net Profit/Year	$3,000,000	$2,550,000	$2,167,500	$1,842,375	$1,566,019
Year over Year Retention Rate	85.00%	85.00%	85.00%	85.00%	85.00%

Retention Program	Year 1	Year 2	Year 3	Year 4	Year 5
Number of Customers	10,000	8,700	7,569	6,585	5,729
Cost of Retention Program/Customer	$5	$5	$5	$5	$5
Total Cost of Retention Program	$50,000	$43,500	$37,845	$32,925	$28,645
Average Profit/Customer/Year	$300	$300	$300	$300	$300
Total Profit	$3,000,000	$2,610,000	$2,270,700	$1,975,509	$1,718,693
Net Profit/Year	$2,950,000	$2,566,500	$2,232,855	$1,942,584	$1,690,048
Year over Year Retention Rate	87.00%	87.00%	87.00%	87.00%	87.00%

Cumulative Differences	Year 1	Year 2	Year 3	Year 4	Year 5
Cumulative Net Profit No Program	$3,000,000	$5,550,000	$7,717,500	$9,559,875	$11,125,894
Cumulative Net Profit With Program	$2,950,000	$5,516,500	$7,749,355	$9,691,939	$11,381,987
Cumulative Incremental Profit with Program	($50,000)	($33,500)	$31,855	$132,064	$256,093
Cumulative Incremental Cost of Retention Program	$50,000	$93,500	$131,345	$164,270	$192,915
Cum ROI	-100%	-36%	24%	80%	133%

Fig. 4.1 Evaluation of retention programs by 5-year ROI

evaluating an overall strategy, one needs to view the impact over longer term horizons. In our example we used a 5-year template. For the more tactical type of projects such as comparing the impact of two communication offers within a campaign, a timeframe of 1 year is more than sufficient. Going back to the CRM retention strategy above, a 5-year time horizon was used since this represents an attempt to keep customers in the long term and not just one year.

From the above spreadsheet, which consisted of a base of 10,000 customers, the company may lose money in the first two years on a cumulative basis but begin to make money in year 3 through the cumulative impact of the increased retention rates from 85% to 87%. As one would expect, all kinds of "what-if"

sensitivity analyses can be done by varying the costs of the retention program. Additionally, a range of improved retention rates around 2% could be considered to determine the overall profit and ROI. Listed below is a sensitivity analysis on a range of improved retention rates around 2% (Fig. 4.2).

Although this example may seem simplistic to produce, it is important to remember that simplicity is essential in conducting these types of analyses. While producing these spreadsheets, it is important to identify metrics or measures that are impacted by the program and, as a result, truly incremental. As seen from the above spreadsheet, "what-if" sensitivity analyses can be very useful in revealing the range of results which could be achieved based on different scenarios. In this example, the results can range from $32 K at 1% improved retention rate to $954 K at 5% improved retention rate. Given this large range of results, it is incumbent that the analyst design a pilot project which produces results with a minimum level of error range. For instance, if the improved retention results are 2% and our desired confidence level is 95%, then the confidence level assuming a sample of 75,000 is 1.9% < = 2% < = 2.1%. One can infer that 95 times out of 100, the expected improved retention rate will be between 1.9% and 2.1% as the program is rolled out. When reconsidering the above sensitivity sheet within this error range, more precise results can be provided to senior management. The profit in this case ranges from lower bound of $233 M to the upper bound of $278 M, which provides far more precision than the original range of $32 M to $954 M and is seen in Fig. 4.3.

Although this example dealt with only one metric (retention rate), which was incremental, in many cases the introduction of a CRM retention strategy can impact more than one metric. In the same example, it is not unreasonable to assume that a retention program might result in an increase of customer spend by $2 each year.

In this scenario, by both impacting retention (2%) and spend ($2 per year), the cumulative profit in year 5 has now been increased from $256 M to $466 M (see Fig. 4.4).

Improved Retention Rate	Cumulative Profit	ROI
1%	$32,966	17%
2%	$256,093	133%
5%	$954,651	466%

Fig. 4.2 5-year CRM ROI at different improved retention rates

Improved Retention Rate	Cumulative Profit	ROI
2%	$233,565	121%
2%	$256,093	133%
2%	$278,669	144%

Fig. 4.3 5-year CRM ROI bounded by 95% confidence level around 2%

Retention Program assuming increase in Customer Value of $2 annually

	Year 1	Year 2	Year 3	Year 4	Year 5
Program	$3,000,000	$5,550,000	$7,717,500	$9,559,875	$11,125,894
With Program	$2,970,000	$5,571,500	$7,849,355	$9,844,833	$11,592,171
Profit with	-$30,000	$21,300	$132,069	$284,958	$466,277
Cost of Retention	$50,000	$93,500	$131,345	$164,270	$192,915
Cum. ROI	-60%	23%	101%	173%	242%

Fig. 4.4 Retention program assuming increase in customer value of $2 annually

A meager increase of $2 per year has almost caused profits to double, which simply reinforces the notion that small numbers can provide huge payback when the results are viewed within a long-term horizon (5 years).

The same kind of thinking can be applied to many other business scenarios where we want to evaluate their overall impact. Another classic example was one that we did to prove whether it was worthwhile to pursue marketing initiatives related to acquisition marketing. The board of a certain non-profit wanted to cancel these types of programs because they lost money in the year 1. Again, we understood the problem and determined the data that we needed to either prove or disprove this claim of the board. Figure 4.5 illustrates this point.

In Fig. 4.5, the data scientists determined the key data elements that were required:

- Average acquisition response rates
- Average acquisition cost per effort
- # of acquisition prospects
- Average value of customer
- Average marketing cost per existing customer
- Average retention rate

The real business problem of understanding the need for acquisition programs then dictated the type of data that was extracted. Once the data was extracted, we then designed the template, which is seen in Fig. 4.5. In designing this, we wanted to provide the perspective that acquisition needs to be examined from a longer time horizon, not just the first year. In other words, we needed to track the program and its results on a year-by-year basis which depended on the required time horizon for payback. In our example here, it was 5 years.

In our example here of Fig. 4.5, we observed that the program paid back in the second year. But the real rationale behind undertaking this exercise was to demonstrate that one needs to evaluate acquisition programs beyond just one year.

Many people might argue that these types of exercises are not good examples of data science in practice. Their arguments would point out that there are no advanced analytics such as machine learning or AI in these examples. But this argument misses the point that data science goes well beyond just the knowledge and application of advanced analytics. In both of these cases, we demonstrate the fact that the data scientist needs to fully understand the business problem and, more importantly, align the right data to solve the business problem.

	Year			
Profits	**1**	**2**	**3**	**4**
# of Customers	1,500	1,050	735	515
Margin/Customer (excluding marketing costs)	$50	$50	$50	$50
Total Margin	$75,000	$52,500	$36,750	$25,725
Costs				
Acquisition Costs	$100,000	$0	$0	$0
Existing Customer Communication Costs/Customer	$0	$3	$3	$3
Total Costs	$100,000	$3,150	$2,205	$1,545
Net Results				
Annual Profit	-$25,000	$49,350	$34,545	$24,180
Cum. Profit	-$25,000	$24,350	$58,895	$83,075
Cum. ROI	-25%	24%	56%	78%
Yr. Over Yr. Retention	**70%**	**70%**	**70%**	**70%**

Fig. 4.5 Using data analytics to evaluate viability of acquisition programs

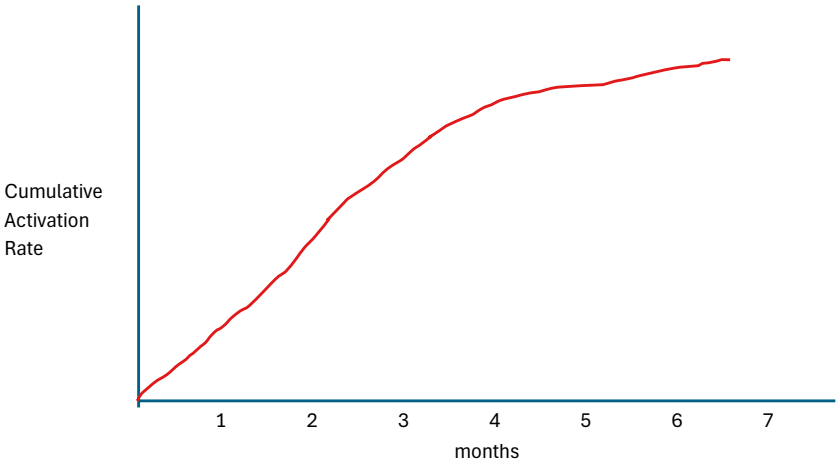


Fig. 4.6 Cumulative activation rate by months of activity

The acquisition and retention examples represent a microcosm of business problems which do not require advanced analytics. Another great example of this type of non-advanced analytics in action were the many times we were asked to determine the average purchase cycle of a customer for a given organization. By simply plotting cumulative activation rates over periods of time (months in this case), we then determined the elbow point which is where the cumulative activation rate begins to flatten. In Fig. 4.6, this occurs at around 4 months.

In later chapters, many of the real-world exercises further reinforce the data scientist's real value in using data to solve business problems of a non-advanced nature.

Key Takeaways

- Skillsets of data scientists can be utilized to conduct business scenario analysis.
- Two of the basic examples that we examined were retention programs and acquisition programs.
- In conducting these business scenario analyses, there are two elements which must be developed by the data scientist:
  - The key information elements to determine success.
  - The design of a template where the above elements are input to determine success or failure.

CHAPTER 5

The Data Science Process-Problem Identification

With Customer Relationship Management (CRM) becoming more of a business philosophy for most present-day organizations, data science is often viewed as the piece of analytical technology which is required to achieve a given solution. But what does this really mean, and does it really describe the true nature of data science? For many people, the fact that data science is viewed as the technological component implies that the purchasing of the right software and hardware is the key to effective data science. Other schools of thought regard the use of statistics and/or computer programming, along with machine-learning-type algorithms, as the data science component. Yet, these preconceived notions miss the mark in the sense that they represent specific components within the overall data science process. This is because data science is a step-by-step process requiring the human element interacting with technology in order to produce the best business solution for a given process. This is best understood by explaining what is involved within each step of this process.

In this book, we consider the data science process as comprising four major steps or stages:

1. Identification of the Business Problem or Challenge
2. Creation of the Analytical File
3. Application of the Appropriate Tools and Technology
4. Implementation and Tracking (Fig. 5.1)

You will get little argument from many experienced practitioners in recognizing that problem identification is probably the most important stage within the entire process. In fact, it is somewhat ironic to hear about the multitude of data science technologies available in the marketplace, and yet realize that this stage of the process is where the human element is most important.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_5

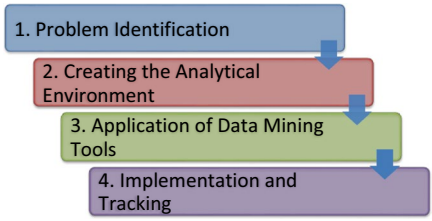


Fig. 5.1 The four stages of data mining

This capability relies on being able to critically assess a given business situation, both quantitatively and qualitatively, and also being able to determine its overall importance given the overall business strategy. This can perhaps best be illustrated by offering an argument:

An organization's overall sales had significantly decreased within the last year. It was also found that 75% of the company's overall sales resulted from their best 20% of customers. The analyst and marketer identified that customer attrition had significantly increased over this period.

Based on this information, it was decided that a defection model to identify high-risk defectors would enable marketers to allocate increased resources to this vulnerable group in the hope of retaining them. Yet this thinking was flawed in two key business fronts. Analysis should have been conducted on the high-value high risk group in order to determine the extent of the defection problem within this segment. Given the preliminary information of this specific business case, it is extremely likely that the defection problem would prove highly prevalent within this segment (high-value/high-risk). The other thing to consider before embarking on the development of a model is to understand whether CRM or data science has any relevance in this situation. For instance, if the reason for losing sales is because the competition has created new products or services which have price points and benefits which are far superior, being superior CRM practitioners through data science is not going to stop this sales erosion. As you can see in this case, some upfront market research is the preferred route to undertake in order to better understand "why" this situation is occurring.

Another important consideration in being able to identify a problem is to understand the current data environment and its implications in resolving a certain business problem. For example, a company may want to launch a new product and develop a cross-sell model in order to target those customers who are most likely to purchase this product. In building a model, the analyst needs to understand that there is no prior information or history on this specific product on which to develop a model. The analyst might then think that there may be other similar products with previous history which could be used to develop a broad profile. If this were not possible, the analyst might want to think of ways to identify customers who could be so-called "early adopters" (i.e. the early pioneer purchasers of new products).

Often enough, the end business user has a perception of exactly what they are trying to derive from data science. In such circumstances, a given solution is etched in their mind without any understanding of the longer-term ramifications. For instance, a credit card company may want to acquire the most cost-effective prospects. From the business end-user's perspective, this may simply involve the development of a response model. Yet from the experienced data scientist's perspective, this request can seem quite vague. The experienced data scientist can quickly point out a number of questions that really attempt to be more specific about the type of solution which should be developed. For instance: do we develop a simple gross response model? This type of solution will yield cost-effective cards in the short term; in the long term, however, the creditworthiness of these new cards will be severely eroded. Although the marketing cost per new card may be dramatically improved, we find that this benefit is more than offset by the increased operations cost of processing these bad new cards. The solution to this dilemma is to build a solution which optimizes both gross response rate (marketing cost per card) and approval rate (operations cost per new card). Yet, a further dilemma may be that there is an extremely high level of attrition within the first six months of acquiring the card. In this case, the business needs to focus on achieving three objectives:

1. Maximize gross response rate with the ultimate objective of optimizing our marketing costs
2. Maximize approval rate with the ultimate objective of optimizing our operations cost
3. Reduce attrition rate with the ultimate objective of increasing the overall first-year level of profitability.

The same kind of example can also be seen in the telecommunication sector, which conducts much of their acquisition efforts through outbound telemarketing. Despite their efforts in attempting to become more efficient at the front end through response models, the business still encounters the negative dynamics of excessive defection within the first three months. Once again, an experienced data scientist would attempt to recommend solutions which provide a balance of acquiring new prospects in a cost-effective manner, yet attempt to optimize the retention rates of these prospects once they become customers.

The tragedy of COVID-19 also brings to mind the importance of problem-solving, and in using the right data in order to arrive at a solution. The obvious goal here was to identify areas where there was increasing incidence of COVID-19. We can all recall the chief health officer appearing daily with trend reports on COVID-19 cases. The reports emphasized cases, and the trend in these cases over time. Most data scientists will quickly identify the alarming flaw here. The data only captured reported cases. It was not capturing all those cases that went unreported, but who were indeed contracting the disease. The biased nature of the data is paramount here. Considering increases in hospital stays does not completely eliminate this bias, as there will still be very sick people who should have gone to the hospital but chose not to. The real

unbiased statistic here is the number of COVID-19 deaths, as there is no so-called "reporting"-type bias here. The obvious limitation with the deaths statistics is that the figures are lagged, but it is nevertheless more robust than the other information which was more prominently displayed. Data accuracy should have been the primary objective, rather than focusing on the lagged nature.

As one can surmise from these above examples, it is evident that some creativity in thinking needs to be employed in ensuring that we are identifying the right problem and challenge given the current business circumstances. Once again, strong collaborative efforts between the marketing or business area and the data science area represent the keys toward really optimizing the creative thinking from both functional areas.

Yet, how does this collaborative effort occur? How can this become a disciplined approach? A request for a meeting with the data analyst (or analysts) is often the first point of contact. The business end-user outlines the problem, and also, in many cases, their vision of what the solution should be. In most cases, this vision is not altogether inappropriate for the challenge at hand. Yet, there are some cases in which this vision is totally inappropriate, or simply does not consider another option which is superior to one that has been deemed appropriate. This is where the data scientist needs to probe and ask more questions in order to obtain as much information and details as possible regarding the business issue. Below there is a list of some of the broad issues or considerations in formulating these questions:

1. What is the overall marketing strategy, and what are the business objectives?
2. What is the specific objective of this initiative?
3. What is the marketing plan, with budgets and forecasts?
4. What historical information and reports do we have?
5. What is the data environment?
6. What are the product dynamics, as well as the competitive/economic forces that will impact the overall results?

What Is the Overall Marketing Strategy, and What Are the Business Objectives?

The overall marketing strategy essentially represents our initial guide in indicating the extent, as well as the complexity, of data science that will be required within a given organization. For instance, the strategy may be one of tremendous growth within a certain sector. The company may tackle this challenge on two fronts. The company may already have existing products or services which are deemed to meet the needs and expectations of customers within this sector. Yet its overall presence within this business sector is extremely limited. On the first front, investment in marketing efforts to increase the company's presence, as well as awareness of its existing products and services, would be one priority.

On the second front, the company would want to investigate new products as well as services to further enhance its overall offerings to customers within this sector.

One good example of this is credit card companies offering insurance services to its customers. Initially, the card may have offered travel insurance or protection in case it was lost. With this customer base, there was also an opportunity to offer other types of insurance, either directly or with business partners. These companies might invest in programs aimed at increasing the awareness of travel and credit card registry insurance. At the same time, funds would be invested in extensive market research to identify the most relevant insurance products and services for credit card customers. Funds would also be invested to test market the research findings as well as to develop an overall launch program. All these efforts serve to meet the strategic imperative of significantly growing the insurance business amongst this customer base in order to become a significant player within this sector. From a data science standpoint, the key focus is significant growth, which implies that costs are not a major priority. Given this objective, companies are willing to sacrifice inefficiencies in order to increase their overall market presence. Data science in this scenario is not a priority, and would be employed sparingly.

However, this will likely change once the initial launch has been completed. As acquisition and growth become ongoing activities, the need for more cost-effective measures will necessitate the use of data science, which becomes most applicable when a given marketing activity is ongoing but costs are increasing at a faster rate than revenues.

Another critical strategy might be the focus on retaining customers. Once again, the data scientist needs to understand that if the focus is on retention, then cost effectiveness may play a significant role. For instance, a given company may have a highly selective group of what they call high-wealth customers. These customers number about 5000 and represent 5% of all customers, yet account for 50% of all revenues. The company may find that attrition rates have increased by 50%. In this case, the data scientist could question whether the identification of high-risk wealth defectors through data science is going to provide real business value. In other words, identifying who might be likely to defect is not as important as understanding why they are defecting. The business owner needs to understand the key motivations/attitudes, as well as competitive forces that are causing large retention problems. As you can see from these examples, probing to better understand the strategy and its objectives provides an initial guide to the overall impact of any data science solution.

In many cases, it is often easy to use data science as a "quick-hit" solution with an obvious immediate net benefit to the company. The demonstration of these results to the executive team in any organization represents an easy avenue for managers to impart their value to the organization. Again, however, these results, despite their net immediate benefit to the company, may actually disguise longer-term business problems that will severely impact the balance sheet in a few years. For example, we may build retention models that allow us

to target high-risk defectors, but which hide the fact that it is overall interest in the company's products and services which is actually declining.

Besides the retention example described above, another good example is the introduction of a new product. Using some basic data science tools, the company is able to achieve an overall net benefit to the marketing program. Yet even with data science, the company's actual profitability has been seriously eroding over the past few years. New competitors have emerged in the marketplace, which has resulted in severe price competitiveness. The company has maintained its pricing strategy and, ultimately, its overall contribution margin on a per unit basis. However, the fixed cost proportion of overall expenses relies on a certain amount of sales volume to meet these expenses. With price competition, the volumes have deteriorated significantly, thereby significantly increasing the fixed cost component as a percentage of revenue. The company's focus at this point should be on the repositioning of the product. Rather than get into the specifics of what this entails, since this is not the purpose of the book, and nor is it my expertise, it is appropriate to state that data science at this stage is clearly imprudent and would allocate resources which should be devoted to other areas of the business.

What Is the Specific Objective of This Initiative?

In many of the initiatives, the solution is relatively tactical, as the marketer or business person already has a preconceived notion of what they want to do. For instance, the development of a cross-sell model represents the type of solution adopted in order to cost-effectively promote other types of products to existing customers. In another situation, the solution may be to lower the costs of acquisition efforts by building response models. On the surface, the decision to build these tools may make perfect business sense, yet one needs to determine if these short-term solutions are consistent with the overall marketing strategy. Once again, if the strategy is to provide significant growth, the use of data science would not be considered as a key component in the initial phases of this strategy. Awareness, rather than cost-effectiveness, is the prime objective of such a strategy. Once this is achieved, the need to maintain awareness as well as continued growth becomes more challenging due to such forces as market saturation and increased competition.

Presuming that the specific objective is consistent with the overall strategy, one may realize that the preconceived solution needs to be thought out. For instance, one could use a cross-sell response model to promote a specific product to the right group of customers. But one might ask "Is this the right product?" Despite having a high predisposition toward responding to an offer for this product, the actual customer may have a higher predisposition toward another product type. The use of product affinity models provides a solution whereby products can be ranked in relative terms of their purchase affinity for a given customer.

In some cases, identifying the business problem or challenge is considered to be predetermined since this occurs before any actual data science takes place. With this approach, one assumes that data science is purely a tactical approach whereby tactical solutions are developed for specific problems and challenges. Organizations that adopt this approach assume that data science either plays a very limited role or no role at all in the development of strategy. It is my belief that this type of approach is sub-optimal. Those organizations which use data science for both strategic and tactical solutions will achieve solutions superior to those which use data science purely for tactical purposes. By including the problem identification component as part of the data science process, we are in fact recognizing that data science can be used to derive insights for the development of new strategies. Let's take a look at a few examples.

A company is beginning to experience significantly eroding sales from their outbound telemarketing acquisition efforts. Their current acquisition strategy has not changed over the last couple of years. Based on this cursory level of information, it would appear that list fatigue has set in and that we need to develop tools that produce more targeted lists. Accordingly, the marketer has instructed the data science department to build a response model to help optimize sales.

But the lack of involvement between the data science department and the marketing department ultimately results in a solution which does not look at the big picture. The acquisition model is built with the intention of maximizing response. As a result, response is maximized but the sales numbers are still sub-optimal. Further investigation reveals that the cancellation rate in the first three months amongst new customers has actually increased. Both marketing and data science made assumptions here that the optimization of acquisition response would lead to optimized sales. If a more upfront and collaborative investigation had been conducted by both marketing and the data science area, then it would have been highly likely that the real problem was to develop a solution that both optimizes response as well as the likelihood to be retained after three months.

Another good example of being reactive (tactical) rather than proactive (strategic) is when the marketing department states that they can't develop a predictive model until we have a marketing database or data mart. Once again, this decision delays the development of tools which can achieve tremendous cost savings for current campaigns. Data science can yield solutions from raw legacy systems and databases in a very short timeframe (some four to six weeks).

Another classic example is the case of "we don't have the data." If the data scientist are allowed to examine the data, their expertise and experience can provide an approach which still provides higher-yielding results. For instance, acquisition programs tend to provide the most varied approaches in how we target names. In today's world, data doesn't seem to be the problem, even in acquiring new customers. Yet data governance issues are now introducing regulatory guidelines on the use of cookies for third-party advertising. This significantly impairs the capability of a given company to better understand the

browsing behavior of a visitor to a third-party website. However, the data could be aggregated at the geographic level, allowing the creation of regional indexes. These geographic areas can then be classified as either higher- or lower-performing based on the browsing behavior of customers in that region. Acquisition programs can subsequently target geographic areas based on the browsing behavior being similar to its desired group of prospects.

Historically, even prior to the internet, Statistics Canada, with all its geo-demographic data, was used to create targeted areas that could be used for targeted acquisition direct marketing programs. This data would contain information such as ethnicity, occupation, education, and a range of other population demographics. It is this rich demographic information that can be used to create geographic indexes that are customized according to the company's targeting objectives.

The actual compilation of this data is at the actual enumeration area level, which is meaningless to all practitioners unless it is mapped back to the more meaningful postal code level. A conversion table allows us to map enumeration areas back to postal codes. Figure 5.2 demonstrates an example template of some of this Stats Can data and how multiple postal codes map to the same enumeration area. In this example, the Stats Can data is unique at the enumeration area. This means that multiple postal codes mapping to the same enumeration area would have identical Stats Can data. These 50,000 distinct points or enumeration areas still provide at least some level of targeting, although data scientists are always seeking more granular type data. A more ideal solution would be if Stats Can compiled their data at the postal code level, since this would represent 800,000 unique points of geographic data.

In any case, however, it is the postal code which is the critical information element used in any geo-demographic analysis.

At an even broader level, Stats Can information was used to carve up the country into 60 PSYTE demographic clusters by one of the truly early Canadian pioneers of data analytics, Environics Analytics.

Although the above issue is Canadian-specific, all countries face similar challenges and issues in dealing with external geographic area data. Difficulties in

Enumeration Area	Postal Code	Median Income	Avg. Age	Avg. Household Size	% of population with university education
100001002	M5A2J1	$42,000.00	40	2	10%
100001002	M5A3J3	$42,000.00	40	2	10%
100001002	M5A4J1	$42,000.00	40	2	10%
100005004	H4B2E5	$50,000.00	35	1	85%
100005004	H4B3E1	$50,000.00	35	1	85%
100006008	L1W3K6	$37,000.00	43	3	5%
100006008	L1W2L5	$37,000.00	43	3	5%
100006008	L1W3K1	$37,000.00	43	3	5%

Fig. 5.2 Example of geographic table (census and postal code) and demographic variables

understanding the granularity of the data or the geographic level of data capture are consistent across all countries. Furthermore, can we append this data back to customer files?

These geo-demographic level models using the above Canadian example will never be as powerful as models developed off individual-level data, but they can still yield very powerful results, especially when applied to list universes that are very large (i.e. 1MM plus). These models in many cases can be more broadly applied since they are developed at a geographic area level. Meanwhile, individual-level models can only be applied against those universes from which the model was originally developed. This can provide limitations when trying to apply the results beyond the confined original universe.

Building models off research data also introduces another bias often referred to as the "responder bias." Not everyone responds to a survey, and it can be a mistake to assume that there are no demographic and behavioral differences between responders and non-responders. One good example of this were the acquisition models developed using market research data from a vast consumer database of 10 million names which was compiled by a Toronto data company. This data company used coupons as an incentive to increase the number of responders. In our analysis of the data, we immediately discovered that there was a high skew toward lower-income females, thereby subjecting our models to bias. Of course, if our target universe was indeed lower-income females, then modeling such individuals would have made sense.

The use of customer penetration indexes, given that there is a reasonable saturation at a particular geographic level, follows the principle of "Fish where the fish are." Using penetration indices (# of customers/# of prospects) within a geographic level as a targeting tool is both pragmatic and easily comprehensible; it is also, more importantly, easy to use. "Deploying these tools does not require a deep understanding of statistics."

In some cases, only very limited information is available to help resolve business problems. Does that mean, however, that data science techniques cannot be employed? Let's explore this more closely through an example. Let's presume that a retail company would like to initiate some CRM discipline within its acquisition programs. As its first order of the day in any CRM initiative, the company would like to use data and information to improve their overall effectiveness. The only information or learning they have is from their research.

This retail company collects no information on its customers. But market research has indicated that the key drivers of purchase behavior are high income, female immigrants.

Here, we can see that the data scientist and marketing team are thinking of creative ways in which to capitalize on existing insights, rather than wait another three to six months to obtain meaningful data. The table below outlines how the creative use of Stats Can information as a proxy for our research learning might provide one targeting approach. Each of the key pieces of market research learning concerning the drivers of purchase behavior represents available information which has been compiled at the enumeration area by Statistics

Canada. Using a conversion table (enumeration area to postal code), we can then create fields of this key information at a postal code level. From this table, we can create an overall index based on this information. Figure 5.3 depicts how this overall index might be calculated.

Assuming that each key field or piece of information is equally important within the overall index, then the index for M5A1J2 is $(.33 \times 1.25) + (.33 \times 1.15) + (.33 \times 2) = 1.45$.

Of course, in adopting this index approach for targetting geographic areas , the appropriate tests and controls would be put in place to evaluate the current approach, but also to develop new learning. This new learning could also be utilized to develop acquisition models if the appropriate random samples are created within the test matrix. Moving forward, we could determine which targetting approach is superior (acquisition model vs. market research derived purchase behavior index). The key point here is that current learning can be actioned immediately, rather than wait for the outcome and results of the next campaign.

Leveraging both the Marketing Expertise from market research as well as Data Science Expertise in identifying problems helps to ensure that the organization is more clearly focused on the higher priorities, which have more significant impact on the business. The marketer's experience, in terms of prior campaign results and strategies, complements the data scientist's experience and expertise in terms of understanding the data environment and what works best for a given business situation. This complementation of skillsets represents a real competitive advantage in being able to identify business problems or challenges that can potentially yield greater returns to the business.

However, it is often the case that the problems or challenges are entirely open-ended. In cases like this, the data science exercise becomes a knowledge or data-discovery exercise which provides direction and strategies on those activities which should be undertaken over the course of the next year or two. This process involves a staged process, with the key stages being:

- Data audit
- Business results gathering and key stakeholder interviews
- KBM: Key Business Measure and post-campaign results
- Segmentation and profiling

We will discuss these stages in more detail in the later chapters of the book.

	Income	% Female	% Landed Immigrant
Average Postal Code	$40,000	52%	5%
M5C 1J2	$50,000	60%	10%
Index	1.25	1.15	2

Fig. 5.3 Creating a postal code (geographic area) level index

Key Takeaways

- This book is written adopting the four-stage approach to data science.
- Identifying the right problem involves close collaboration with the business stakeholders.
- Business problems can be both strategic and tactical, with strategic problems requiring a data-discovery exercise, which is discussed in a later chapter.
- Acquisition models often yield poorer results than existing customer models, but the benefits can be greater as they are deployed against millions of records.
- Existing customer data can be aggregated to a geographic area and then used as independent variables in acquisition modeling.
- Data scientists need to be creative in using their skills when there is, seemingly, no solution.

CHAPTER 6

The Data Science Process: Creation of the Analytical File

Once a business challenge or problem is identified, analysts must understand the data and information requirements necessary to conduct meaningful analytics. This doesn't require the same rigorous data needs analysis that is involved in building or designing a database; rather, the analyst focuses on the data that already exists, not on what should exist. Once a file and its contents are identified as potentially relevant, analysts typically request the entire file as one data component in the project.

Additional files may be requested depending on their relevance and whether customer-related information can be extracted. At this stage, the data audit process begins. This rigorous exercise aims to optimize the data environment for analytics by focusing on:

- Data quality
- Data integrity
- Usefulness for analysis

When source files are identified, analysts must evaluate data quality. For example, certain fields may have a high proportion of missing values.

In Fig. 6.1, 43% of the customer base lacks a start date. Techniques such as imputing average or median values can handle missing data. More robust methods may involve building predictive models based on other database fields. Another poor-quality variable example is when all records share the same outcome, like gender in Major League Baseball analysis—which is useless for comparison due to the absence of female players.

Data integrity issues arise when field values don't make sense. In Fig. 6.2, most product codes consist of letters, likely referring to product categories. However, a category such as "999" warrants further investigation.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_6

Tenure	# of Customers	% of Customers
1998	49,000	14%
1999	49,000	14%
2000	59,500	17%
2001	42,000	12%
Missing	150,500	43%
Total	350,000	100%

Fig. 6.1 Frequency distribution of customer tenure

Product Category Code	# of Customers	% of Customers
ABC	103,810	30%
DEF	118,650	34%
GHI	74,165	21%
999	49,875	14%
Total	350,000	100%

Fig. 6.2 Frequency distribution of product category code

After addressing data quality and integrity, analysts consider how to group or summarize fields. For example, purchase history may be summarized annually or by total lifetime spending. With hundreds of product codes, grouping them into broader categories allows for meaningful future analysis. Yet this execution must be done in order to ensure that the grouped information has enough data for any future statistical exercise.

Once these tasks are completed, algorithms organize the data into one overall analytical file. This is where analysts add significant value by creating derived variables, such as trends or time-based purchase patterns, which are not directly present in the source files. In fact, roughly 90% of variables in a typical analysis are derived.

Merging multiple files into a unified analytical file is critical and requires understanding pre-existing match keys or logic to create them. Analysts must recognize relationships between data files: one-to-one, many-to-one, or many-to-many. Data can be categorized as:

- Behavioral (e.g., customer actions)
- Non-behavioral (e.g., demographics, geographic data)

Ultimately, analysts must determine which variables are useful for the analysis at hand, alongside the record of interest.

All data practitioners are universally in agreement that real data science success resides in the quality of data. Although the use and knowledge of statistics is important, the lack of decent or reliable data will result in sub-optimal solutions. Intimate knowledge of the data environment represents the real

"golden nugget" towards producing effective solutions. "Intimacy" may sound like an odd word when dealing with data, but it does express a certain level of closeness that one needs to obtain when conducting any data science exercise. This feeling of closeness implies that the analyst needs to truly understand the reliability and integrity of each piece of data that is being considered for a given data science project.

The example of the education process can best serve to illustrate the importance of data. In the education system, the two core disciplines of reading and mathematics are considered to be essential requirements in any successful career. Yet in both these disciplines, there are basic technical fundamentals that need to be mastered before one can be successful in either of these disciplines.

In the area of mathematics, one needs to master the fundamentals of addition, subtraction, multiplication, and division. In reading, the student needs to master the alphabet or ABC's before he or she can even attempt to read a story.

This type of process for mastering a particular discipline is no different than the process for attempting to master the world of analytics. In the world of analytics our ABC's is **DATA, DATA, and DATA**. Analytics practitioners must not only have a conceptual understanding of data, but also a hands-on approach towards data. This approach affords the analyst a much deeper understanding of the data, as they are then better able to handle all the detailed nuances which are a fact of life in the data world.

Conducting a Data Audit

In any environment, whether in the digital or the offline arena, the analyst needs to have a process of better understanding data. How is this done? The first step is the data audit, which will be the focus of this discussion.

Once the data sources have been identified for a given project, they need to be mapped onto the source data arising from the source files within their IT environment. The appropriate source data is then extracted, whereby a data audit is conducted on this data. The extent or detail of this data audit would depend upon whether or not the project is in development mode or simply reproducing these reports on an ongoing basis.

If this is the initial creation of these reports, the source data and source files are likely to be unknown to the analysts. In this case, the initial development of these reports would require that the more exhaustive data audit be performed in order to explore each field or variable from all the source data. These reports would provide insights on the following:

- Missing values
- Value ranges
- Minimum/maximum/average
- Number of unique values

The results from the data audit would indicate the usefulness of a given variable within the overall set of measurement reports. If, for instance, age is important, but 75% of its values are missing, its usefulness is limited. However, a binary variable indicating missing vs. non-missing age values could still offer insights.

Regardless of whether or not the data represents new information that is unfamiliar to the analyst, or even if it represents a familiar existing data source, a process such as a data audit process needs to take place that allows a better understanding of the data at a detailed level. An initial step in this process is viewing a sample of 10 records (a "data dump") to ensure data loads correctly and to get a feel for its complexity (Fig. 6.3).

In this above example, something has gone awry as the third record reveals values which do not make sense for all the fields that are to the right of the birth date. Further investigation reveals that missing values were not handled properly when the data were loaded into the company's analytical environment. Fixing this problem reveals what is shown in Fig. 6.4.

Looking at the number of files, as well as the number of fields on each file along with examples of the values of these fields, the analyst now begins the quest towards a very detailed understanding of the data environment.

After this initial glimpse, further data diagnostics are conducted which allow the analyst to determine the number of unique values and the number of missing values for each field on each file that is received by the analyst. Figure 6.5 gives an example of this kind of report.

In the figure, household size appears unusable with 90% missing data, while product type, due to its character format and with 3000 unique stock keeping unit (SKUs), must be converted into binary variables (1 = presence, 0 = absence). With only 0.03% presence per SKU, such variables are ineffective for modeling; this is often referred to as the issue of high cardinality.

Further diagnostics (e.g., frequency distributions) help understand value distributions (Fig. 6.6).

In the above report, this might indicate that it would be difficult to establish personal communication with these people (50% of customers) since we do not know where they live.

In Fig. 6.7, gender might be a difficult variable to use analytically as 50% of the variable values are missing. However, as stated previously, we could create a variable based on reported gender. Yet this issue of missing values for income would be easily addressed by simply replacing the missing values with

Account Number	Postal Code	Birth Date	Start Date	Behavior Score	Income	# in House
123456	M5A 3S6	Jul-49	Mar-91	500	30000	6
345321	H3A 2B5	Aug-54	Apr-92	550	42500	1
543235	T5A 2S7	Jun-83	600	35,500	3	543210
etc.						

Fig. 6.3 Sample of three customer records and fields

Account Number	Postal Code	Birth Date	Start Date	Behavior Score	Income	# in House
123456	M5A 3S6	Jul-49	Mar-91	500	30000	6
345321	H3A 2B5	Aug-54	Apr-92	550	42500	1
543235	T5A 2S7		Jun-83	600	3500	3
etc.						

Fig. 6.4 Sample of three customer records and fields, adjusted

Variable	# of Records	Data Field Format	# of Unique Values	# of Missing Values
Income	100,000	Numeric	50,000	2,000
Customer Type	100,000	Character	4	10,000
Gender	100,000	Character	3	50,000
Household Size	100,000	Numeric	7	90,000
Product Type	100,000	Character	3,000	5,000
Customer Name	100,000	Character	100,000	0
Postal Code	100,000	Character	50,000	0

Fig. 6.5 Data diagnostics report

either the mean or the median. Figure 6.8 is another example based on tenure (creation date).

Note here that we observe a significant increase in customer counts between 2005 and 2006, indicating that acquisition in this period increased significantly. This would warrant investigation into why this might be occurring.

Creating an analytical file also involves defining pre- and post-periods. The post-period contains the objective function (e.g., reducing defection), while the pre-period includes data to predict that outcome.

With the information from these data audit reports, the analyst can then begin to determine how to create the analytical file, which essentially represents key information in the form of analytical variables. The objective in this exercise is to have all meaningful information at the appropriate record level where any proposed solution will be actioned. In most cases, this represents the customer or individual, but it is not necessarily confined to that level. For example, pricing models for auto insurance are actioned at the vehicle level while retail analytics might focus on decisions that are actioned at the store level.

The data audit process should be conducted not only on new information or data sources but also on known data sources that serve as updated or refreshed information for a given analytical solution. Although the data audit process used in processing a new data source can be quite comprehensive, a less rigorous process may be used when assessing refreshed or updated data. This may be as simple as checking the number of records and fields that are received, as well as some basic data audit reports that look at means or averages of key variables which are unique for a given client. In any event, the purpose of even doing a shortened version of this data audit report is to establish a means of identifying problems or issues with the data that is currently being used by the

Fig. 6.6 Frequency distribution of region

Region	# of Customers	% of Total Customers
Europe	25,000	2.5%
North America	100,000	10.0%
Africa	350,000	35.0%
Asia	25,000	2.5%
Missing Values	500,000	50.0%

Fig. 6.7 Frequency distribution of gender and income

Gender	% of Records
Male	23%
Female	27%
Missing	50%
Income	**% of Records**
<25000	25%
25000-50000	25%
50000-75000	25%
75000+	23%
Missing	2%

analyst. Figure 6.9 gives a template example of what such a report might look like.

In this above report, key variables or potentially Key Business Measures (KBM's) are looked at over time to identify potential data issues or changes (both positive and negative) that are taking place in the business.

Data audits, despite producing the least glamorous type of reporting information, are necessary prerequisites in any analytics process. If one believes that all analysis starts with data, then one must exercise extreme diligence and respect for the data. This kind of attitude towards data helps to foster an appreciation of the many kinds of data nuances that can appear in a project. If analytics is going to be successful, data audits represent the first critical task within any given project.

With this initial solid understanding of the data, the other critical component in creating the analytical file is the concept of pre-/post-periods. In any data science exercise, we need to understand the metric being analyzed or the metric being targeted or modeled. This information is often referred to as the objective function or the behavior we are trying to better understand. Here are some common business examples below:

- How do we reduce attrition?
- How do we reduce credit loss?
- How do we optimize response?

In all these above examples, the analyst is constructing the analytical file in such a way as to create a solution to the above questions. But the behaviors or objective functions that are being analyzed ideally should be created in what is

Column Name: CREATIONDATE	Count	Percent	Cumulative Count	Cumulative Percent
[Missing]	1,822	3.4%	1,822	3.4%
2002	2,276	4.3%	4,098	7.7%
2003	2,027	3.8%	6,125	11.5%
2004	2,430	4.6%	8,555	16.1%
2005	3,213	6.0%	11,768	22.1%
2006	5,575	10.5%	17,343	32.6%
2007	6,699	12.6%	24,042	45.2%
2008	7,793	14.6%	31,835	59.8%
2009	7,239	13.6%	39,074	73.4%
2010	8,743	16.4%	47,817	89.8%
2011	5,418	10.2%	53,235	100.0%
Total	53,235	100.0%		

Fig. 6.8 Frequency distribution report on creation date (tenure)

Fig. 6.9 Historical perspective of key business measures

	Period 1	Period 2	Period 3
Record Count			
Avg. Purchase Amount			
Average Age			
Average Tenure			
Etc.			

referred to as a post-period. Only the information used to create the objective function should be used in a post-period.

Meanwhile, the pre-period is used to create all the information which will be used to derive insights about the objective function or behavior in the post-period. A schematic (Fig. 6.10) best illustrates how we might look at this.

In this schematic, the analytical file is actually constructed so that all the pre-period information (red) occurs prior to the information in the post-period. Hence, in a way, we are using the analytics in a more robust way by looking at information (pre-) that could be interpreted as predictive of something that will happen in the future (post-).

As practitioners, we often assume that the objective function or post-period variable is relatively straightforward and easy to create. Certainly the old adage of "never assuming" once again applies here.

Many examples exist in which the dependent variable or objective function is not straightforward. The first example is the use of proxies in creating a dependent variable. Requests are often submitted to build predictive models where no prior history exists. If there is no prior history, we cannot build a model that precisely predicts this type of behavior, since it has not occurred. However, we can look for historical behavior that is similar to that which we want to predict. This "similar type of behavior" becomes our proxy for building the model. For example, we may want to build a model that predicts the likelihood of someone purchasing "tennis shoes", as the company has decided to promote these types of products. A reasonable proxy might be the

Pre period	Post Period
All behavior prior to Jan 01 -2011	Jan 01 - 2011 / Mar 31 - 2011
All predictor variables	Objective function

Fig. 6.10 Schematic of pre- and post-period on the analytical file

likelihood of someone purchasing a tennis racquet, where these products have been promoted in the past.

Another good example might be the purchase of squash products, since none of these products have been promoted in the past. Under this scenario, do we use tennis products as a proxy, or do we consider the notion at a much broader level? Using tennis racquet purchase behavior could certainly be considered as a reasonable proxy for tennis shoe purchases. Does this still apply, however, when considering a completely different sport? Some analytics may be required here to give us an insight. Correlation and affinity analysis may be conducted to find out where product purchase behavior is more closely aligned. If we discovered that individual sports, such as running, golf, and tennis, were more closely aligned than team sports such as baseball, hockey, and football, then we may want to identify our objective function as someone who is most likely to purchase a product related to an individual sport.

Building models to predict car purchase behavior can be a very challenging endeavor, as such purchases are very infrequent, with a purchase cycle of around five years. The notion of trying to build repurchase models becomes daunting, since some customers will purchase a new car at three years, some at four years, some at five years, and the rest at six or more years. Certainly, analytics could be conducted to determine the time period of most of these purchases. In any case, we are probably looking at a purchase period that encompasses a number of years as our post-period in defining the objective function. The creation of the analytical file would look at all the characteristics and behaviors of the customer prior to the post-period. If the post-period is three years, then we are using information which is arguably somewhat dated. A better approach to narrow down this post-period window might be to use lease information, whereby we specifically know when the lease expires. We could model off these lease expirations and determine who renewed and who did not. This information could then be used as a proxy for car purchase, since it looks at customers who are engaged with the car company. Using engagement by modeling off lease renewal represents a more practical approach towards determining car purchase behavior.

Retention models often present the most difficult problems when trying to create the dependent variable. In most cases, there is no precise behavior that defines cancellation, such as renewing a subscription. Rather, the determination of cancel behavior is defined by looking at inactivity within a specific period. As we observed in the car example above, however, defining this period is not straightforward. For example, customer purchase inactivity of a week may define someone exhibiting cancel behavior within a grocery company,

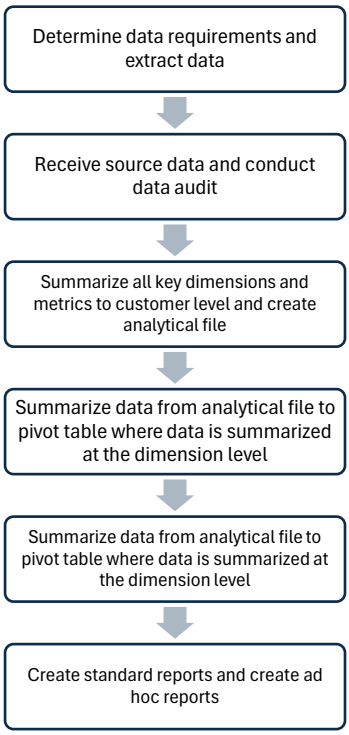


Fig. 6.11 Flow chart for creating business reports from pivot tables

whereas using the same period of inactivity to define cancels for tire purchases would be inappropriate. Once again, analytics can be conducted to more precisely define what the purchase period should be.

Pivot Tables and Measurement Reporting

Analytical files may be summarized into pivot tables. The complexity of this summarization will increase as the number of dimensions and the range of outcomes within a given dimension increase. For example, if the average spend for a given business initiative is assessed in terms of gender (male and female) and education level (college-educated and non-college-educated), then there are four (2 × 2) possible views or dimensions, which are often referred to as cubes. However, suppose that spend is analyzed by gender, education level, age category (<25, 26–35, 36–45, 46–54, 55+), and household size category (1, 2, 3, 4, 5+), then the number of views (or cubes) by which spending may be analyzed, explodes to 100 (2 × 2 × 5 × 5) possible views (or cubes). The pivot table itself represents the source information in terms of creating the desired measurement reports. Figure 6.11 is a schematic of what this process might look like from the original data sources to the creation of the measurement reports.

The explosion of digital data has simply augmented the need for increased reporting or business intelligence. Although pivot tables can facilitate the development of customized business reports, there are limits to their use. For example, there is a limit of 1 mm rows in Excel, which limits the reporting capacity as the data becomes more granular. Think of two variables, which each have 1000 unique values. The number of possible views (or cubes) here is one million. If we want to look at any other variable, we easily exceed the one million row limitation. This is the reason why visualization companies such as Tableau and PowerBI have entered the marketplace. With their software technology, users have access to an infinite number of views (or cubes), resulting in these software companies being hugely successful within the analytics industry. Planning what to measure and understanding the data deeply are both essential for meaningful reporting. Collaboration between marketers and data analysts ensures success.

Case Study: Retailer Example

One retailer for which we worked was seeking to develop a CRM (Customer Relationship Management). It began with the data audit and then developed into best customer program. The objectives of this exercise were twofold:

1. To understand the data environment
2. With this understanding, to develop a best customer program.

For objective 1, the standard data audit program was conducted. Besides the dumping of 10 records, diagnostics were also created, which indicated the total number of records in each file that were loaded into the system. This was done to ensure that whatever was received from I/T was identical to what was loaded into the system.

Once this data was loaded into the system, standard data audit diagnostics were completed on all files. A frequency distribution is listed with the following insights (Fig. 6.12).

This above report revealed that 50% of customers had no known address. In other words, any personalized contact to these customers which required the use of address would be extremely limited. Of those customers with a known location, 70% resided in Ontario. Both these findings provided learning that: (a) direct marketing or any CRM type system would be severely limited with

Fig. 6.12 Frequency distribution by region

Region	# of Customers	% of Customers
Prairie Provinces	25M	2.50%
Quebec	100M	10%
Ontario	350M	35%
West	25M	2.50%
[Missing Values]	500M	50%
Total	1MM	100%

Last Name	First Name	Address	Postal code	Phone #	Customer ID
Boire	Richard	123 Main Street	L1W3K6	905-489-2124	1000123
Boire	Richard	123 Main Street	L1W3K6	905-489-2124	1000126
Boire	Richard	123 Main Street	L1W3K6	905-489-2124	1000129
Boire	Richard	123 Main Street	L1W3K6	905-489-2124	1000123

Fig. 6.13 Example of same customer mapping to separate customer IDs

50% of the customers having a missing address and/or postal code; and (b) most activity seems to be occurring in Ontario.

Another component of the data audit exercise was to match customer files from the different retail stores in order to determine the quality and integrity of the "supposed match key". The understanding was that customer number, or ID, was the unique qualifier for each customer. In fact, not only was this customer ID unique; it was also too unique. Primary forensics involved matching the customer file to the billing file. Here, the usual outcome was produced in that one customer ID matched to many customer IDs in the transaction file. Of course, it is an expectation to find this outcome, as customers will have multiple transactions with a given company, indicating that there is a one-to-many relationship. Yet, further forensic investigation on this file, when looking at dumps of records, revealed the following scenario. Figure 6.13 shows what my record looked like.

This outcome indicated that a given phone number had multiple customer IDs. Now, of course, one could argue that perhaps there may be more than one person in a household with the same phone number that are actually separate customers. However, deeper investigation of the data was conducted by looking at the name and address of those records, and it was discovered that duplicate phone numbers did not indeed map to separate customers, but rather to the same customer (Richard Boire example above matches to four unique customer IDs). In these cases, the actual same person still mapped to multiple IDs. In presenting this information to the company, it was decided that a better understanding of the process of how a customer obtains a given ID was required. In the investigation, it was discovered that a customer obtains a new ID if it is his or her first time in each store. This led us to the discovery that customer IDs were not unique to a given person; rather, they were unique to the person and store. In other words, if a person conducted business at five different stores, they would have five separate IDs. In the example above, Richard Boire had gone to three separate stores (1000123, 1000126, and 1000129) and had shopped twice in the same store (1000123).

In thinking about how to proceed with any further analysis, it was realized that phone number would be the optimum match key. Each store was always required to capture the phone number every time the customer made a purchase. Specific recommendations included the following:

1. Use phone number as the basis to build unique customer keys.
2. Use software which can append name and address to customers with missing values in these fields. With phone number available on 100% of the records, this is the critical key in a table provided by the software that contains fields which maps phone number to name and address.
3. Begin to consider programs which expand the retailer's presence beyond Ontario.

The key consideration here is that the better insight to the existing data environment allowed us to identify and adopt these above data strategies.

The second component of the research dealt with being able to conduct some type of analysis. In discussion with this retailer, it was discovered that they had no knowledge of who their most valuable customers were. Were 80% of revenues being driven by 20% of the customers, or was it more like a 60:40 ratio? The basic Pareto rule regarding customers and revenue contribution was completely unbeknownst to them. Yet, even in determining how revenues distributed amongst customers, the company wanted to better understand what this looked like and to ultimately create a profile of their best customers. With this objective in mind, data was extracted that would provide the necessary solution. The extraction focused on active customers only (i.e. customers that had purchase activity in the last year). This led to the first key finding: half their customer base was inactive, which represented a significant future business challenge, but one that was outside the scope of this initial project. Figure 6.14 is a decile table in which active customers (those who had had at least one transaction in the last 12 months) are ranked by total purchase amount.

Note that purchase amount was agreed to by the retailer and is being used as a proxy for value.

The second objective was identifying the best customers. Analysis focused on active customers from the past year and the major findings are listed below:

- 50% of customers were inactive.
- 60% of revenue came from 20% of customers.

Segment	% of Customer	Avg. Annual Sales Last Year	% Contribution to Total Sales by Interval
High Value	0 - 5%	250	25%
High Value	5 - 10%	170	17%
High Value	10 - 15%	110	11%
High Value	15 - 20%	70	7%
Low Value	20 - 25%	50	5%
Low Value	25 - 100%	25	37%

Fig. 6.14 Example of value segmentation report for retailer

Product Type	% of All Customers	% of All High-Value Customers
Bought Product A	20%	40%
Bought Product B	30%	10%
Bought Product C	25%	25%
Bought Product D	25%	25%
Total	**100%**	**100%**

Region	# of Customers	% of All High-Value Customers within Reported Postal
Prairie Provinces	5%	5%
Quebec	20%	30%
Ontario	70%	55%
West	5%	10%
Total	**100%**	**100%**

Fig. 6.15 Examples of high-value profile reports

Customer Segment	Control	Test	Do Not Promote
High Value/ Best Customer	175,000	5,000	5,000
Balance	5,000	5,000	5,000

Fig. 6.16 Example of simple test matrix for best customer program

Consensus coalesced around the notion of creating a best customer program for this top 20% of customers (high-value) group. Yet, our work was not finished, as marketing also wanted to better understand what these customers looked like (Fig. 6.15).

The best customers tended to live in Quebec and the West and preferred Product A. A campaign targeting the 200,000 best customers was launched.

Looking at where these customers came from, and what products they purchased in the past, provided a cursory-level profile.

- Best customers tend to reside in Quebec and the West
- Best customers tend to buy Product A

With this information, the next step was to action learning within a "best customer" program. A list of names for this program needed to be created. Figure 6.16 is a schematic of what this list selection looked like.

Note how the names (175,000are all frontloaded as the segment (Best Customer) and communication strategy (Control) are expected to be the winning strategy. The remaining cells are about learning:

- Did the campaign yield incremental purchase results over and above those who received no campaign offer (best customer control vs. best customer do not promote and test balance v. do not promote balance?
- Which communication strategy works best (test vs. control): high value control vs. high value test and balance control vs. balance tes?

- Do higher-value segments (best customer) yield more incremental results than lower-value segments (balance) (high value control vs. balance control and high value test vs. balance test)?

The program was launched and easily exceeded their expectations, thereby justifying the shift from mass marketing to CRM marketing.

Key Takeaways

- The data audit and its three key reports are a critical tool in understanding what will be useful for future data analysis
- Missing values and cardinality are key elements which need to be assessed during the data audit
- The dependent variable or objective function for a predictive model is often not readily available but must be derived based on other information.
- The record of interest is the input record which will go into any modeling algorithm or reporting routine
- The case study from one of our previous clients clearly demonstrates that the analytics behind a "best customer" program comprised two phases:
  - A data audit.
  - Identification of best customers and what they looked like.

CHAPTER 7

Data Science Process-Creation of the Analytical File: Using External Data Sources

Today, virtually all organizations which conduct activities in data science utilize external data sources, which in some cases is purchased, but for the most part is readily available on the web. These data sources represent information which a given company is unable to collect based on its internal activities. Another perspective on this is that this type of data represents information which is available to the public either free or at some cost. The level of data science sophistication will determine the extent of these external data source purchases. The reasoning behind this type of data is to augment existing customer-level information. But the question is: how can it augment data at the individual customer level, especially given the challenges of matching individual customers to these external data sources? Yet, most organizations will use geo-demographic data, which is available from Statistics Canada. Despite its breadth of information, its primary limitation is that all demographics and behaviors are aggregated to a postal area level. Let's further examine the value of this aggregated information. The availability of variables such as age, income, and gender is one primary data-gathering objective of the data scientist. Yet, it is not uncommon to observe demographic variables at an individual level, such as age, income, and gender, with more than 50% of their values absent. External data sources, including aggregate geo-demographic data, can fill this void.

Using External Data for Existing Customer Programs

First of all, organizations must determine whether the data needs are for acquisition programs or for existing customer programs. In the case of the latter, most organizations in today's CRM-driven environment have, at a minimum, both offline and online databases of existing customers and their purchases. The organization and structure of this data may vary from company to company, depending on its level of database marketing sophistication. Yet,

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_7

regardless of how this data is structured and organized, for data science purposes, the data is at an individual level. As discussed above, data scientists will always strive to use as much individual-level data as possible, since this will always provide superior results. However, these results can be compromised if there are quality issues with the individual-level data. For example, age and income might be reported on an individual level. Yet if over 90% of customer records contain missing values in these fields, this individual-level data on the remaining 10% is not going to very useful in a data science project. In this case, one should look at aggregate data sources, such as Statistics Canada, specifically, Stats Can Census data, or the appropriate external data provider within that country. Aggregate-level data means that customers residing in the same postal area would have the same Stats Can values while customers residing in different postal areas would have different Stats Can values. Appending aggregate-level data (Stats Can Census area) to this customer file would at least provide full age and income information to 90% of records which do not have any of this information. The use of this aggregate-level data for income and age would yield superior results to the status quo individual-level data, where 90% of the information is missing.

Besides enhancing information where there are data quality issues, aggregate-level data can also provide breadth of information. For example, categories such as ethnicity, occupation, religion, education, and a range of other Stats Can demographic information are unlikely to be directly available on any customer database. Appending this above types of information to an existing customer database can add some value to a modeling or profiling exercise. However, it is the rich individual-level information related to the purchase or transaction information of the customer that will produce the stronger modeling/profiling variables, which are relevant for targeting purposes. This aggregate-level demographic information can provide broad insights that can be used to develop better communication strategies. For example, the demographic and geographic profile of a certain group of customers may indicate that these customers live predominantly in areas with high income, a high number of immigrants, and a large percentage of self-employed people. Given this information, marketing teams might want to develop unique communication strategies around these groups, rather than promote the standard generic benefits of the given product.

The bigger impact of external data will stem from its use in acquisition programs. In fact, data suppliers more often market themselves on their overall impact in acquiring new customers. This need for external data is much more significant in building any data science solution for acquisition programs than in providing supplemental information for existing customer programs. Typically, name, address, and postal code represent the available pieces of data for any acquisition program. Postal code, for example, is the key link in being able to append Stats Can data to name and address records. Stats Can data is offered under two main types of products. The first product is Stats Can tax filer data. The data here is organized at a postal walk level, representing

approximately 800 households and contains information compiled from annual tax returns. Such data would contain income-related information, including income earned from employment, investment income, charitable deductions, etc. The second file, Stats Can Census data, contains information at an enumeration area level (or approximately 400 to 500 households). Although it contains some income data, it lacks the other measures of wealth that are contained in the tax filer data. However, this file is much richer in terms of demographic information and contains information pertaining to ethnicity, religion, language, education, occupation, etc. Furthermore, it is much more granular, as data is aggregated for every 400 households as opposed to 800 households, as is the case with the tax filer data. Another limitation to this data is that it is only updated every 5 years, the period between Stats Can Census surveys.

What to Consider When Purchasing External Data for Data Science Purposes

Our current data environment has certainly mitigated limitations regarding the lack of data. Despite its aggregate nature, there is often the need for certain accurate demographic data. Purchase decisions regarding these data sources can become extremely important in any data science exercise. In purchasing data for data science purposes, the data is not used to replace existing data, but rather as additional data sources to the existing data environment. But how do we determine the value of this data, and can we observe that this new data does indeed provide significantly better results over what currently is within the existing data environment? The use of lift charts, which will be discussed at much greater length throughout the remainder of the book, represents one such option. For example, the incremental benefit is the additional rank ordering or incremental lift provided by the external data if used within a predictive model or targeting solution (see Fig. 7.1).

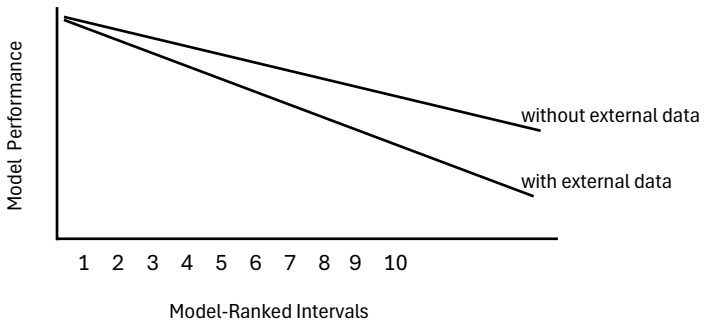


Fig. 7.1 Lift charts with and without external data

Key Takeaways

- Data is much more readily available in today's digital data-rich environment.
- Primary external data sources that are most commonly used are Stats Can Data.
- Despite the aggregate nature of Stats Can geo-demographic information, its breadth and coverage make it a valuable data product.

CHAPTER 8

Data Storage and Security

It is often thought that the largest reason for the prominence of data science in today's business world is software, and the specific mathematical/statistical applications that are embedded in these tools. However, many of these tools have been in existence for over 25 years. The underlying reason for the historical lack of use within the business world has been expense. The costs involved in providing the appropriate computer resources to run these applications were simply exorbitant. In today's technical environment, the limitation is not necessarily on the costs of processing, but rather on the ability of the human being to ingest this information into something meaningful. In fact, the use of MapReduce technology from organizations such as Apache Hadoop has even accelerated this capability to a new level. The world of Big Data has given rise to technology that allows us to identify the Five V's of Big Data: Volume, Variety, Velocity, Value, and Veracity (see Fig. 8.1).

Large amounts of data (Volume) with differing formats (Variety) are being streamed in real time (Velocity).

These technological advances have led to the wider use of data science techniques because of two factors:

- Increased ability to store and access data
- Increased ability to process instructions

The increased ability to store, access, and process data is fundamentally a function of the significant advancements and reduced costs of both computer memory and disk drive storage technologies over time. The improved availability and affordability of computer memory (RAM) facilitates the temporary storage of much larger data sets, and allows more and faster processing (reads, writes, calculations) to be conducted against data that resides in memory, compared to data that resides on permanent storage devices (disk drives). The

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_8

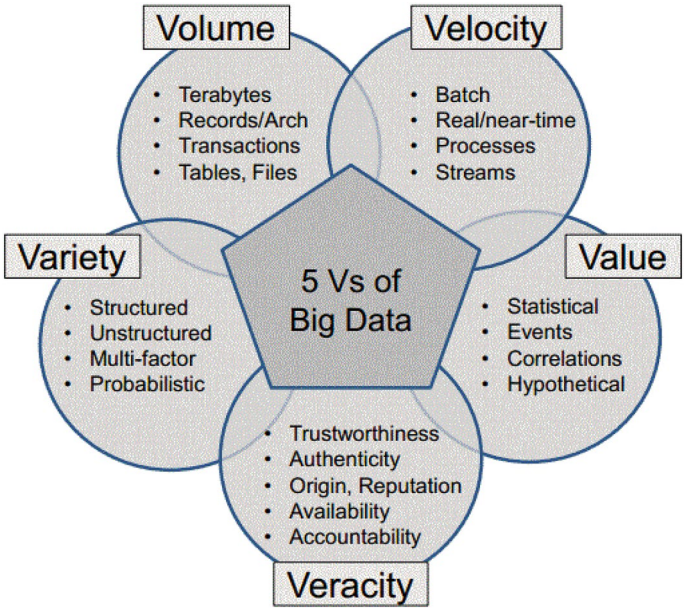


Fig. 8.1 The 5V's of data

improvements in permanent storage devices (disk drives) have enabled the online storage of much larger quantities of data at reduced costs. This has provided faster read and write capability and greater reliability, allowing much greater amounts of data to be stored in an "online" (readily accessible) state. From a business standpoint, this has translated into the ability to access large amounts of data at ever-increasing speeds.

As data represents the primary tool of the data scientist, the security of this data or information, as outlined in legislation, is of acute interest to the data scientist. The protection of data that is entrusted to data scientists is not an option, but rather an obligation. Data scientists are constantly exposed to the transfer of data either in receiving or sending data. During this transfer of data, the data scientist should certainly be aware of the sensitivity of the information. This sensitivity will dictate the various protocols and procedures which are required in any data transfer. Expectations regarding the handling of data have changed considerably over time. Currently, sensitivities regarding data theft and loss have contributed to more stringent use of protocols regarding the transfer of data. Encryption techniques and secure file transfer protocols (SFTPs) are the norm in today's environment.

Another mechanism for ensuring that data is in a secure environment is to have proper storage procedures. Data scientists, as a rule, would like to keep data as long as possible and have it readily accessible to them. But the reality is that there is always a certain degree of risk in keeping data for extended periods of time, especially data containing names and addresses. Because data scientists

deal with data that is both used for creating and applying a solution, they also need this data when validating the solution that has been applied to a campaign. For instance, in analyzing the results of a direct marketing campaign, the name and address information is used to distinguish responders from nonresponders within the campaign. The rule of thumb, though, in retaining name and address is usually six months, which means that the expectation is that a backend or performance analysis should be done within six months of the campaign launch. If files are older than six months, procedures should be in place to archive this data to offline storage.

In today's digital environment, data security is even more paramount given the data governance procedures that are now standard for most medium to large organizations. Yet, this has simply increased the need for more sophisticated encryption techniques, as well as the need for more powerful firewalls. In today's Big Data world, large organizations have leaned towards the creation of a data lake, which is basically a repository of all data, both offline and online. Here, the data is unorganized and essentially in its raw source form. Further organization of this data by the organization would involve both data engineers and data scientists as they attempt to organize this data in a more meaningful manner.

But the need for processing and accessing this data is critical, as the essence of success for any AI solution is extremely large volumes of data. Think of the success of various generative AI (GENAI) tools such as CHATGPT. Their tools access all the data which is on the internet. Their investors have contributed large sums of money in order to provide them with the right tools and platforms to harness this data access. Currently, the most formidable challenge facing humankind is climate change, as these tools consume tremendous amounts of energy, thereby creating additional pressures on existing initiatives. Existing research is now exploring avenues that still create meaningful solutions, but which limit the degree of impact.

The explosive growth of data will continue. As data science practitioners, we should be aware of the data governance protocols from both governments and organizations. Additionally, given the current data environment of almost unlimited data storage, we need to be cognizant of tools that provide access to data, while not conflicting with our efforts on addressing climate change.

Key Takeaways

- Increasing technology has improved access to much larger volumes with a variety of different formats.
- With AI, and particularly GENAI, access to huge volumes of data is a must.
- Yet this increased access has resulted in higher energy requirements, thereby yielding negative impacts on our efforts to improve climate change.

CHAPTER 9

Privacy Concerns Regarding the Use of Data

Privacy is a hot-button topic about which people will have strong opinions. In Canada, much of this opinion has been formulated into many of the principles and guidelines as outlined in the Personal Information Protection and Electronics Document Act (PIPEDA). In many cases, the legislation is very clear about what businesses can and cannot do. However, there are shades of grey when terms such as "reasonableness" are used. For the most part, businesses are highly respectful of the legal framework since the legislation itself has been a best practice by many organizations over a large number of years. It is simply good business practice to be respectful and attentive to the privacy needs of consumers when attempting to properly market the right services and products to them.

In many ways, the practices and disciplines being exercised by organizations will result in legislation that could either be very tolerant or very restrictive. The legislative direction within this area will ultimately be determined by the extent to which businesses are respectful with regard consumers' rights to privacy. This implies that data scientists, as well as business stakeholders, be proactive rather than reactive regarding the privacy issues that they encounter on a daily basis.

Although there are ten guiding principles within legislation, there are three that have the greatest significance for the area of data science:

- Security or safeguards
- Identifying purposes or use of information
- Consent
- The principle of security or safeguards was discussed at length in the previous chapter, so here we will focus on the first two principles here in this chapter.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_9

- In many ways, the use of data science on customer data is not as privacy-contentious as might be expressed by some of the leading privacy experts. Why? The use of mathematics and science essentially allows business users to make decisions on groups and not individuals. The reality of any data science analysis is that business users will apply data science results at an individual level, yet recognizing that the insights and learning regarding these decisions are based on the results and performance of a group of individuals.

How Will You Use My Data?

Within the actual consent clause, the text should be clear to the purpose of the consent. For example, if the purpose is to use the information to either rent or sell a given name to a third party for marketing purposes, then this activity should be clearly stated in the clause. In the digital world, the leading technical companies, such as Google and Facebook, use consumer browsing behavior to identify target audiences for a given company's ads. Most consumers are unaware that this information is being used to target ads to them, yet wonder why they receive all these ads on their digital devices. Neither Google nor Facebook have been overt in their communication to consumers that they will use the user's browsing behavior and sell this information back to advertising agencies and organizations. The ability to monetize this browsing data has allowed both Facebook and Google to become multibillion dollar-type enterprises without the consumer being totally aware about the use of their data.

In today's digital world, consumers should be made aware that their digital browsing behavior will be used to generate insights and to potentially target them for future digital promotions. But how much detail should the company provide in terms of how they generate these insights? For example, there have been some opinions expressing the notion that the consumer be informed that mathematical and/or machine learning techniques, including AI, will be used to help analyze their information.

But should businesses really have to explain the intricacies of AI and machine learning, as well as other mathematical techniques, to the public at large? This was certainly one issue which was discussed between myself, a data scientist practitioner, and other privacy stakeholders. The issue here like, as with most aspects of life, was "trying to do the right thing." But what is "right" when dealing with consumer privacy? But how does one know what is right? In many cases with regards to the legislation, "right" attempts to look at what might be term "reasonableness." That is the crux of the matter. Is it reasonable for consumers to know the arcane details of how businesses are analyzing their data? Do consumers really care that data science tools and statistics are used to analyze their data? If these tools are used to select consumers, then it might be argued that the consumer cares not about how they are being selected, but rather that they might be selected for particular services and products. The use of these tools to help in selecting a certain individual is of no interest to that

consumer as long as they understand that they could potentially be selected for a particular offer or promotion. Today, this might be the threshold of reasonableness for the consumer. But, as stated earlier, as more experience is gained within the application of the law, the threshold of reasonableness might change to the point that consumers need to be informed about how this information will be analyzed. If that is the case, then the book *Statistics for Dummies* will become a best-seller.

Consumers also expect that future marketing activities which would be using this information are somehow related to the initial marketing activity in which they gave their consent. For example, it is reasonable to expect that a credit card customer, having giving consent, could be offered insurance to protect his overall credit balance in case of a job loss. Yet, it is unreasonable for a donor of a non-profit company to expect to receive a promotion for a new credit card.

This example should not to be taken to indicate that non-profit organizations are restricted in their activities. It is meant to imply that all organizations, which also include "for-profit" organizations, need to consider what is reasonable from the customer viewpoint. Are the offers and services of the organization somehow related to each other so that the customer can understand why they are receiving the promotion in the first place?

Reasonableness is a grey area, and certainly the test of reasonableness is implicit throughout the legislation. This is why many pundits believe that this legislation is constantly evolving, especially in today's digital information age. This is already evident by the European legislation Genertal Data Protection Regulation (GDPR), which was created in 2018, as well as legislation in California and other states. In Canada, ongoing efforts are being undertaken to update the PIPEDA given our current digital environment.

Consent

Consent can appear to be relatively straightforward in terms of whether or not the customer gave permission to use his or her information. However, this principle is often more complex since it ties in very closely with how a given company will use the information. For instance, there is a great difference between a company promoting products and services to its existing customers and the use of information which can then be sold to a third party. One might argue that the former activity, of continued marketing promotion to existing customers requires less restrictive means of obtaining customer consent, than the latter activity of monetizing information for use by third parties, which is a huge revenue generator for companies like Facebook and Google. In fact, selling customer information versus using customer information would suggest that different levels of consent are required. These different levels typically involve two options: opt-in vs. opt-out, where opt-in is the more stringent option. Let's discuss these two options in more detail.

Opt-In vs. Opt-Out

The notion of opt-in vs. opt-out type consent represents a very important distinction in terms of gaining customer consent. In opt-in-type consent, the customer or individual must initiate some activity before customer consent is deemed to be obtained. This is often in the form of a checkbox within some type of offer or communication piece, which is being promoted by the company. The customer must fill in the check box before customer consent is considered to be given. In other cases, the use of a questionnaire to obtain additional customer information might contain a clause, stating that filling in the questionnaire is the recognition of implied consent to use both questionnaire and behavioral information for future business purposes.

From a marketing standpoint, the opt-in for obtaining customer consent is very restrictive for most organizations and would severely impact their ability to conduct their business in a profitable manner. Many people will do nothing when it comes to receiving information about proactively asking for consent. But it does not necessarily mean that they do not want to receive additional products or services, or even that they care that their information is being used to better understand their browsing behavior. It could simply mean that they have not paid attention to the opt-in check box. Yet, the use of an opt-in type clause is extremely restrictive in that very small volumes of customers would proactively give permission to be either promoted or to use their information. The net eligible customer universe available in marketing offers and products would be significantly smaller, such that many marketers would not be able to take advantage of the economies of scale within a much larger volume universe. Many of these companies would cease activities that had previously been very profitable and simultaneously provided value and service to the consuming public. Because these activities are no longer profitable, jobs would be lost, along with their consuming dollars, which, of course, would be detrimental to our economy. The use of this information would also be very limiting as it is primarily based on the notion that those people who are opting-in may not be representative of the population that we are trying to analyze.

Because of the deleterious effect of opt-in clauses, many organizations have attempted to deal with obtaining consent using the so-called opt-out provision. In the opt-out provision, the customer must actually fill in a check box in order to state that consent is "*not given*". A blank response to this check box is recognized as implied customer consent. The key for companies which choose to use opt-out consent is that this consent opt-out clause should be clearly visible and prominently displayed within the overall text. For many new companies that are developing the latest digital apps for users, user consent is implied or opt-out. This means that the user gives consent to the use of any information by downloading the app. And there is a clause in the user agreement, which stipulates the user's implied consent. Typically, however, the clause is buried toward the end of the agreement.

One might argue that this type of requested consent by these new digital organization is, at the very least, misleading to the user. Any attempt to bury the consent in small print and within the middle or end of the text defeats the purpose of the overall privacy legislation. Yet, opt-out clauses that are very clear and upfront to the user represent those best practices for those organizations that pride themselves as marketing experts and which help to further cement the relationship of the customer with the organization.

Channel Consideration

In customer consent, another key consideration is the type of vehicle or channel that is being used to promote products and services. Do direct mail, text messaging, programmatic advertising, email, and outbound telemarketing all require the same level of customer consent? This is a question which is, at present, subject to impassioned debate. Some channels are perceived to be more intrusive than others. For example, receiving a direct mail piece is certainly going to be less intrusive than receiving a phone call just after supper time. As organizations and governments continue to place increasing emphasis on data governance, the issue of channels has not been a significant consideration in the overall discussion. Perhaps, different thresholds should be used which are dependent on the channel; I expect this will be the norm as data privacy becomes more engrained as a core business philosophy.

Time Period for Consumer Consent Decisions

Suppose a consumer chooses not to give consent to future promotions and/or offers. Another one of the areas which the legislation does not address is the period of time under which the consumer's decision should be honored. Should a name forever not be marketed once he or she has refused consent? Marketers realize that this "do not promote" file will grow over time and that, at some point it will represent an opportunity. This is somewhat comparable to what happens with customers that lapse within an organization. Although a customer has done nothing for a lengthy period, the marketer will, at some point, attempt to resuscitate the activity of this lapsed customer. For "do not promote" names, the logic works in the same way in that, after some period of time, it would appear reasonable to attempt to reassess the status of these customers. Perhaps after a year, a promotion specifically geared to reassess the status of these customers might be tried. An opt-in statement could also be presented in their monthly billing statement if they ever decided to change their mind.

The same type of issue also arises when a person gives consent. What is a reasonable time period for customer consent? As with the non-consent issue, there is no real hard rule. Certainly, the legislation doesn't address the issue of period of time that is relevant for both consent and non-consent. Suffice it to say that this is one of those issues that need to be addressed in an increasingly data-driven world.

What about the difference between industry sectors? Should there be different levels of data governance by sector? Certainly, one sector, the health sector, given the heightened sensitivity of patient data, is already dealing with this through its specific Health Insurance Portability and Accountability Act (HIPAA) legislation.

All these above examples in this chapter refer to customers or patients, but what about prospects? Suppose a company sources its names by scraping the web as part of its overall acquisition strategy and decides to build a prospect history database. As the database becomes populated, they begin to run acquisition campaigns against this information. Over time, they can collect and store information related to type of promotion, frequency of promotion and recency of promotion at an individual prospect level. This information is tremendously valuable as organizations can segment prospects based on promotion frequency and recency. Under this scenario, are organizations being respectful of consumers' right to privacy? One could argue it both ways. Certainly, there is no way that the consumer would know that this type of information is being used to target them. But let's consider the actual information that is being used in more detail. The actual information that is collected relates to what the company has done to the prospect in terms of previous promotions. This type of information is company-driven rather than consumer-driven. There is no consumer-driven information here which relates to specific consumer activities or consumer transactional behavior. From this perspective, this prospect history information could arguably be considered somewhat less sensitive since no consumer-driven information is contained on the database. Given its less sensitive nature, marketers might be inclined to use this information to better target their prospects without being overly intrusive.

Yet, in a world of ever-increasing digital marketing, we are all being bombarded by messages from organizations with which we have no relationship. As stated above, the use of our digital browsing behavior has become a highly valued monetizable asset, as seen by organizations such as Facebook and Google. GENAI (Generative Artificial Intelligence) tools, such as CHATGPT, all use our digital data in producing their solutions. Might someone ask Sam Altman (CEO of CHAPGPT) if there are procedures in place to respect consumer privacy as they scour the web for all the data that needs to be used in building their powerful solutions. The simple answer is that this GENAI technology is moving so fast where most people have virtually no knowledge of GENAI and AI, or at least its basic mathematical underpinnings. As these technologies continue to evolve, parameters and regulations need to be instituted to provide the necessary guardrails that will eliminate abuses and optimize benefits. One definite area is data governance and privacy.

Using Credit Risk Data

I have already briefly discussed the sensitivity of data in the health industry. Another contentious sector is the use of credit risk data. This has always provided excellent information in attempting to target customers for products that

take into account the customer's ability to pay. Certainly, the consumer expects that this information is going to be used in making credit decisions since this is clearly stated in any application where the consumer is seeking credit.

However, in many marketing models, it was often one of the strongest variables. However, you might ask whether the consumer reasonably expects that this information is going to be used for marketing purposes. Clearly, the test of reasonability would fail here and, in fact, the use of credit risk data for marketing purposes is simply not a current business practice. However, organizations like Equifax, which understand the value of data, have created data products that contain this credit-related information, but at a postal code level or zip level. Organizations purchase this data as another source of geographic-level information which can be overlaid onto the database. Although the geo-level or postal code/zip code credit risk data is not as powerful as credit-risk data at an individual customer level, the credit risk information at a postal area level can still provide another piece of valuable information which can be used to help target customers for various products and services.

Before I close this chapter, let us consider one well-publicized case which was perhaps the most seemingly public breach of data privacy: the Target case of the pregnant teenager. In this case, the analytics team at Target had identified this young woman as being someone most likely in need of baby products. However, this young woman, who was pregnant and single, was still living at home with her parents. When the household received a targeted mailing about certain baby products, this was opened by the father of the young woman. He was shocked that his young daughter was receiving literature related to baby products, since he was unaware that she was pregnant. Nevertheless, the public relations challenges were enormous for Target when this was reported in the press. One might argue whether Target is really at fault here. After all, the daughter was a customer of Target with an address that just happened to be the same as that of her parents. Is it reasonable to expect that Target, in conducting campaigns to hundreds of thousands of customers should have known that the daughter's maternal situation was unknown to the father, especially since the communication was directed to the daughter and not the father? But then again, given the sensitivity of the product, maybe Target should have ensured that the women receiving these promotions were either married or living by themselves. Yet, as they say, hindsight is 20/20.

Data privacy and governance have become significant areas of concern in our present-day data-driven world. Yet continued advances in technology along with increasing data proliferation would suggest that the privacy advocates are always in a state of "catch-up" mode. We need to recognize that no one is a true real expert in this area since consumer privacy and governance is an ongoing and evolving concern. The lawyers, though well versed in the discipline of law, do not have the deep understanding of data science/machine learning and analytics that industry experts do. As a result, the data people are dealing with legislation which has been designed to protect the consumer, but without the

lawyers necessarily understanding the practical business implications as well as the data and technology used to solve these problems.

But rapid changes in data and technology would suggest that the regulators and government need to enact a more nimble process in ensuring that any advances are adhering to the principles of data privacy and data governance. Lawyers need to have a deeper understanding of data but not trying to turn them into data scientists; at the same time, data scientists need to have a deeper understanding of its laws without them becoming lawyers.

Key Takeaways

- Data Proliferation through social media and AI are making it more encumbent for organizations to be compliant with data privacy
- The Canadian legislation (PIPEDA) has 10 main principles with three overarching goals:
  - Did I give you consent
  - How will you use my data
  - How will you protect my data
- There are two consent mechanisms with option being more restrictive than opt out. All digital apps use the less restrictive opt-out type of consent
- Data Science stakeholders need to be more cognizant of privacy rights while lawyers and legislators need to have better understanding of data and its benefit to society.

CHAPTER 10

Types and Quality of Data

Differences Between Data Levels

If there is one opinion of data science about which almost all experts are in agreement, it is that the data components and content are the core of any data science success. Despite the latest advancements in technology and software, which all purport to significantly improving data science results, data is clearly the driver behind any successful data science project. Having said that, it is important to understand that there are significant types and levels of data. Knowledge of these types and levels of data can provide tremendous insight into obtaining an overall sense of how performance will be impacted.

Data scientists tend to look at the quality of data with regards to four main areas:

- Data granularity
- Biasness of data
- Data coverage
- Range of data values

Data Granularity

In any data science exercise, the objective is to acquire as much individual-level data as possible. Since solutions will always be applied to individuals, development of a solution at this level allows the analyst to utilize unique values of a field for each record. Let's take a look at an example (Fig. 10.1).

What are some preliminary observations that can be identified regarding these six customers without experience in statistics?

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_10

Fig. 10.1 Example of customer/individual-level demographics

Customer	Age	Income	Response
1	75	80K	Yes
2	28	40K	No
3	45	60K	No
4	33	45K	No
5	62	70K	Yes
6	65	75K	Yes

- Older customers are more likely to respond
- There is a relationship between age and income, where higher age translates into higher income

Now suppose geo- or aggregate-level data (namely Statistics Canada demographics at the census level. Bear in mind the same notion can be applied to many other countries as demographic data would also be collected at the appropriate geographic level for that country) was utilized. Typically, this consists of aggregate-level data at an enumeration area level containing information such as income, age, gender, education, ethnicity, occupation, language, mobility, etc.

This information is typically appended to the customer file by postal code using a conversion table containing both postal code and enumeration area. Customers residing in different enumeration areas will have different Stats Can values. However, customers residing in the same enumeration area, but in different postal codes will have identical Stats Can Census values. In appending data, the data scientist will always attempt to obtain the most granular type of data. This implies that data summarization to a given level is minimized. For example, appended data could be available at a variety of levels:

- FSA Level: approx. 10,000 households
- Postal Walk Level: approx. 800 households
- Census/Enumeration Area: approx. 400 households

From the above, it is obvious that the census data would be the most attractive option given our objective of granularity. However, the postal walk data contains financial information, which is not available at the census level. This includes data such as:

- Charitable donations
- Registered Retirement Savings Product (RRSP) room and contribution
- Dividend and investment income

Furthermore, postal walk information is also updated every year whereas the census data is updated only every five years. Depending on the objective of the data science project, different purchase decisions will be made regarding postal

Census Level	Age	Income	Response
10010001	75	80K	Yes
10010001	28	40K	No
10010001	45	60K	No
10080002	33	45K	No
10080002	62	70K	Yes
10080002	65	75K	Yes

Fig. 10.2 Example of six customer-level records with census-level data

Census Level	Age	Income	Response
10010001	49.3	60K	0.33
10080002	53.3	63.3K	0.66

Fig. 10.3 Example of six customer-level records rolled up to census level

walk vs. census data. For countries outside Canada, the tradeoff between loss of granularity and relevant information for a given project would still apply.

To truly appreciate the loss of information when using external aggregate-level data, let's consider how this data might appear if the individual level data existed for six customers on age, income, and response along with the census level. Listed below is the table for these six customers (Fig. 10.2).

In Fig. 10.2, the trends are quite clear in that responders are much older and have higher incomes.

Let's say that now this information is only available at the census level. Here is how this information is summarized (Fig. 10.3).

By comparing both tables, we can see that our findings regarding trends differ when we consider each table separately. For instance, the above findings at the individual level are not as pronounced when summarized at a census level. The relationship between age and response is minimal, while the relationship between income and response is also relatively weak.

These simplistic examples illustrate the point of how information is lost when aggregating data to some level. The notion of keeping data at its most granular level will always be a paramount objective for data scientists.

Data Bias

One of the fundamental questions that need to be asked at the beginning of any data science exercise is whether the data is representative of how the solution is going to be applied. What is meant by this? Let's take a look at an example of an organization that had been hugely successful in data enhancement services prior to 2000. They conducted huge consumer surveys from which the results and learning could be applied to the Canadian population. It was discovered that the respondents of this survey were overwhelmingly female

(approx. 80%). Applying any learning from this sample would, obviously, be erroneous since the usual population gender split is about 50/50. Of course, one of the unique skills of good data scientists is being able to work with the data they have in order to arrive at some solution. In this case, the data scientist might have stratified the above consumer survey sample to obtain an analytical sample that does indeed have a 50/50 gender split. This would involve the random selection of every 4th female survey responder to be included in the survey.

Another helpful example is the application of solutions developed at a national level that are applied regionally. Without getting into any political debate on the issue, it is understood that Quebec differs from the rest of the country. This is primarily because of the fact that the majority language in that province is French, whereas English is the majority language in the rest of the country. Seasoned marketers will state that what works in the rest of the country does not work in Quebec. This suggests that solutions there need to be developed solely from information collected within that province. In fact, it has also been my experience that applying solutions developed in Quebec to French-speaking citizens outside that province are, indeed, not usually optimal. A similar phenomenon is observed when applying Rest of Canada (ROC) solutions to English-speaking Quebec residents. In other countries, the same regional disparities need to be considered when applying data science solutions. For example: are the solutions that are appropriate in Texas also applicable in New England?

These two examples illustrate that some basic demographics need to be considered when trying to evaluate whether or not there is a bias in the data. Besides region and gender, other characteristics to consider are income, age, and household size. In trying to determine the biasness of a given sample, a simple spreadsheet exercise can be produced that displays the basic statistics (means) between the Canadian population and the sample being considered for the data science project. Let's look at an example (Fig. 10.4).

Demographics	Canadian Population	Sample for Data Mining Project
Live in West	20%	19%
Live in Quebec	28%	27%
Live in Ontario	40%	41%
Live in Maritimes	12%	13%
Income	$38,000	$37,000
Age	45	44
% Female	51%	49%
# in household	1.5	3

Fig. 10.4 Comparison of population vs. sample averages

From these results, it is evident that the sample has a large bias toward larger families. Applying any solution from this sample would imply that its impact would be greatest on larger families.

These types of bias observed are not necessarily bad if the biased nature of the universe is understood before it is applied. It is also important to understand how different or biased the data environment is between the time of developing the solution and the time of its application. By comprehending this difference, it is possible to select various alternative courses of action, such as reducing the performance expectation of a given solution or developing a new more optimal, but specific solution for this biased universe.

Data Coverage

Although files or databases contain many fields, this does not mean that the information is useful. It is not uncommon to observe database environments with many fields, but with few recorded values in any of these fields. Dealing with missing values, or data fields where most of the records have no recorded value, is one of the most common tasks that a data scientist must undertake. The use of a data audit (previously discussed) allows the analyst to uncover how prevalent the missing value situation is for every given data field. With the use of frequency distributions, it is possible to obtain a very good sense of the usefulness of a variable based on the extent of missing values. One useful rule of thumb is that any variable or data field with more than 90% of the records having no recorded value would not be considered a worthy variable in any future data science analysis. Organizations can use the results of this component of the data audit to help create a standard list of variables which would serve as a template for all future data science exercises.

Range of Data Values

Examination of the range of values for variables allows the data scientist to identify potential outlier issues, particularly for variables that are continuous in nature. Determining the number of unique values also provides insights into what the data scientist may need to do in terms of manipulating the data. For example, if the entire analytical file is represented only by males, there would only be one value for gender. In this case, gender would be a meaningless variable in any future data science exercise because there is no other outcome which can be analyzed for comparative purposes. In some cases, character type variables such as product codes may comprise hundreds of different outcomes. This kind of scenario would indicate to the data scientist that some grouping or categorization of these outcomes needs to occur rather than just creating binary variables for each outcome. An even better example of this is how data is analyzed geographically. Analysis by geography tends to be done at the region or province level. An analysis of postal codes or zip codes does not provide enough mass in terms of who lives in the postal/zip code (very few)

relative to those who do not live in the postal/zip code sample (everybody else in the sample).

Data Enhancement Services

The business of supplying data has grown significantly over the past several decades. This growth can be attributed to the growth in data science.

Historically, clusters were used solely to target prospects. As the level of sophistication has increased regarding the use of data, however, analysts are looking not only at the clusters but also at the raw source data which was used in building these clusters. The analytical mindset is to consider a variety of data inputs when building data science solutions. With this kind of mindset, the data suppliers themselves have become more sophisticated by offering data enhancement products beyond just cluster codes. One company offers data enhancement products which provide increased targeting capabilities amongst the different ethnic groups. Another organization offers demographic postal area information that goes beyond the information provided by Stats Can census and tax filer data. In fact, this organization uses both of these sources, as well as a variety of other data sources in being able to provide annual demographic and behavioral data at a postal code level. Although this data is much more granular (postal code level), and would therefore appear to be superior to the more aggregate-level Stats Can census or postal walk data, it is important to know that this postal code demographic data represent estimates and not raw or source data. Estimates are derived using the mathematics of nearest neighbor modelling and, because they are estimates, they are going to have some degree of error. Yet, in building data science solutions to differentiate customers and/or prospects based on a desired behavior, the use of this type of data can indeed provide more powerful inputs to a data science solution than simply using the raw Stats Can data. But again, data scientists will always strive to consider using both the derived granular estimates as well as the more aggregate raw source Stats Can data as data inputs in any solution. However, it is extremely important that the data scientist understand what source data is actual vs what is estimated when building a given solution. Furthermore, an understanding of the level of granularity (i.e. postal code vs. enumeration area vs. postal walk vs. Forward Sortation Area (FSA)) provides additional insight into how a given solution might perform. As stated before, more granular level-type records yield, in theory, better-performing data science solutions.

In the business-to-consumer (B2C) aspect of marketing prior to the growth of the internet and digital data, some businesses provided data enhancement services where demographic and behavioral information were collected at the individual consumer level. The data was collected through massive direct mail surveys which were distributed on a quarterly basis. This created a massive database of 10 million Canadians with individual demographic and behavioral information.

The internet and its growth have facilitated the use of external data by companies in its capacity to provide individual browsing behavior. The speed of this medium in its ability to capture external data is a clear competitive advantage toward companies that are collecting data through the slow process of direct mail surveys.

In the business-to-business (B2B) world, though, data enhancement services is still a required service given its capacity to provide firmographic information. Some of this firmographic data relates to the following:

- Type of industry
- Years in business
- # of employees
- Annual sales
- Head office vs. branch
- Key contact

In Canada, there are almost 5 million businesses (4.74 million as of 2025). This presents huge marketing opportunities in using firmographic information to target businesses for various products and services.

More About Missing Values

The transformation of missing values, as we have noted earlier, is a critical process within data science. Imputing an average for missing values of continuous variables, or using a default value for missing values of categorical variables, is acceptable in many situations. But a more exhaustive and robust approach might be considered for these variables that are known to be highly significant within a typical data science process. For example, certain insurance products have higher appeal depending on the age of the customer. In some cases, age might be the strongest predictor of response to this insurance product. But we may only have age data on 50% of the customers. Imputing an average for the remaining 50% would reduce its ultimate impact on any targeted behavior as we are diminishing the impact of any trend based on the continuous nature of the variable. Using predictive modeling techniques or some type of Chi-Square Activation Interaction Detection (CHAID) analysis, we can impute or estimate the age of the missing value records based on other information. Although, the age is not a "hard" observed age but rather an estimate, it is still better to have estimates that more closely resemble the continuous nature rather than impute one value, such as an average, to the entire missing value group.

Key Takeaways

- Data granularity is the goal for many data scientists when using data for a given project due to its potential in building superior solutions.
- Data coverage is another goal for data scientists as they seek data which is very broad-based and not restrictive in nature.

- Biasness of data is a key consideration in building any solution. Approaches can be accommodated to mitigate against this bias.
- Stats Can still remains a key source of external data for most organizations in Canada.
- The internet and digital data now provide tremendous opportunities as external data sources.

CHAPTER 11

Segmentation

In attempting to initially understand customers, the first step should be to segment them. Although consultants will all agree on this first step, they will differ over the approach to be adopted when doing so. The differences in segmentation reside in the fact that there are practical options versus scientific options. Rather than simply choosing to adopt one approach over the other, the correct choice will depend on the complexity of the given customer base, as well as the nature of the current business challenge. For instance, a bank's customer database containing 1 million customers may require a much more sophisticated approach than the retail customer database, which contains 500M(thousand) customers. Identifying students as a customer segment for a student program within a bank is going to be far less complex than trying to identify unique customer segments, where the objective is to create distinct corporate strategies around these corporate segments.

In adopting a course of action for segmentation, it is important to remember the old business adage of the "KISS" principle: Keep It Simple, Stupid. This has tremendous relevance for the following reasons:

- It is easier to understand
- It is easier to execute ease in tracking

In developing any segmentation system, the incremental benefits need to be weighed against the complexity. Complexity will increase the challenge in achieving a broader understanding of what it means throughout the organization, as well as the ability to both execute the system and track its results.

The first consideration should be the number of segments to be created in the database. Some organizations have their customer database segmented into more than 50 segments. In adopting or developing an initial segmentation system, a good rule of thumb is to consider dividing into no more than 10

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_11

segments. Keeping the number of segments to a minimum means that initial learning can be acquired, which may provide insights that support either increasing or decreasing the current number of segments.

Some critics may argue that this suggestion is too broad in terms of identifying the right number of customers for a particular program. However, the idea of a segmentation system is to develop several broad segments with specific models that could be applied to each system.

As suggested, segmentation can mean different things to different people. In one case, it may simply be using business rules such as selecting all customers who have been a customer for longer than two years, who have an income of more than $100 K, and who have bought more than two products. In another case, it may be scoring customers through a model, or perhaps a value-type metric, and then selecting these customers based on score. Another scenario may involve the use of statistics to determine distinct or homogenous groups of customers. No one approach is necessarily superior to any other. They are all good approaches; it is simply that one may be more appropriate than another, depending on the business problem and, more importantly, the data. Both simple and complex solutions can be acceptable to a particular business stakeholder. Similarly, not all segmentation requires sophisticated statistics. In some cases, a simple and more pragmatic approach will suffice. So, how do you determine whether to use the simple approach or the complex approach in any particular context?

Simple vs. Complex Approach

The first consideration when choosing between the simple and the complex approaches depends on the data environment. Why? Simply put, more data implies higher-quality analysis and more purposeful application to marketing activities. Yet this will depend on the quality of the data itself as well as the variety. More complex solutions will arise from these types of data-rich environments.

The second consideration concerns the volume of customers. A larger volume of customers offers more potential to derive significantly larger dollar benefits, even when the lift becomes more marginal. Consider the situation of a 1% business lift on 2 million customers versus a 5% lift on 200M(thousand) customers. From an absolute standpoint, the larger opportunity lies with the larger volume of customers, despite its much lower lift potential. With this larger volume of customers, more complex solutions can be explored.

Let's look at an example to bring some of these scenarios to life. Suppose a company sells one product and collects the billing information of the customer for the last two years as well as their name and address. The current number of customers is 100M(thousand). Before moving forward, the company needs to determine how this information is going to be utilized. More specifically, is this information going to be used for targeting or for communication purposes? Of course, most people would say both. Yet any specific solution will only be able

Fig. 11.1 Example of two-cluster solution on top 50% of customer billers

	Tenure Average	% with pre-authorization plan
Cluster 1	2.5 years	40%
Cluster 2	6 years	20%

to resolve one of the above purposes. For instance, if targeting is the prime objective, clustering is not required. One could merely rank-order names based on the desired business objective which, in this case, is billing amount or sales. Through rank-ordering, customers are placed into groups or deciles where the top decile represents the highest-billed customers, and the bottom decile contains the lowest-billed customers. Groups of customers can then be selected for a marketing initiative based solely on their prior total sales (billing) with the company. However, what if the marketer wants to establish more meaningful communication with this company's customers? The next question would be: what other information besides sales can be used to help in this process? In the example, there is no opportunity to use other information for communication purposes since no other information beyond the billing amount exists.

However, suppose this same company collects payment types and keeps track of a given customer's billing history. In such circumstances, it is now possible to create two other key variables. The first one, tenure, is developed using the customer's first billed date as a proxy for tenure. Meanwhile, payment type can be used to separate customers into pre-authorized payment and non-pre--authorized payment. Along with this other newfound information, the target group of customers may be identified as being the top 50% based on their total customer billings in the past year. Now, it needs to be determined whether this tenure and payment-type information is rich enough to develop unique communication programs. Indeed, a simple clustering exercise might help to decide this in a quantitative manner by scientifically determining whether or not there are some distinct customer groups based on tenure and pre-authorization plan to these top billers. The data and results might look as in Fig. 11.1.

In this simple example above, the clustering exercise does provide additional insights. Here the marketer can develop a communication program to newer customers who are more likely to pay by automatic withdrawal, which would differ from the communication program needed for longer-tenured customers who pay by cheque or credit card.

Looking at the Complex Example

In the above example, it can be determined whether or not tenure or pre--authorization plan has an impact on being a top biller. If either or both characteristics have an impact on being a top biller, then communication programs could be built around the top 50% without undertaking any clustering. Yet, if profiling is the more reasonable and practical option in this case, one may ask why the company does not segment all customers based on value and then

simply profile the high-value customers from regular customers, as opposed to just developing clusters around the high-value group. The answer, in many cases, is that this is an acceptable solution. In cases of companies with large customer volumes (i.e. much larger than 100M(thousand) customers) and very rich data, however, this may not be the optimal solution. For example, a bank may have 5MM(million) customers where the top 2.5 million customers are deemed as having high enough value to be considered as the net eligible group for future CRM initiatives.

Profiling these top 2.5 mm against the bottom 2.5 mm reveals six key characteristics that best differentiate both groups:

- High tenure
- Live in Toronto
- Have a mortgage
- Have multiple investments
- Have multiple loans
- Have Visa
- Have high transaction fees

This information can be deemed a waste of time by rightfully arguing that this is a profile of a high-value customer which could have been created without doing any data science. So, the question remains about how this information is used to develop unique communication programs. It is already known that the target group is 2.5 mm customers. But how does a marketer talk to them? This is where clustering plays a hugely significant role. Through clustering, the analysis could provide these types of clusters (Fig. 11.2).

This information shows that three separate programs (note the numbers highlighted in red) could be developed to high-value customers. By comparing the variable means between clusters, we can identify certain variables for a given cluster where their variable means are different than the other clusters' variable means. For cluster 1, programs would be developed that perhaps communicate the advantages of different types of long-term lending schemes (mortgages) as well as financial benefits that are more advantageous to someone living in a

	Cluster 1 Average	Cluster 2 Average	Cluster 3 Average
High Tenure	3	10	3.5
Live in Toronto	0.8	0.5	0.45
Have a Mortgage	0.9	0.4	0.49
Have Multi Investments	0.3	0.35	0.7
Have Multi Loans	0.4	0.5	0.8
Have Visa	0.6	0.9	0.65
Have High Transaction Fees	0.5	0.8	0.55

Fig. 11.2 Example of three-cluster solution for bank

large city like Toronto. In the case of cluster 2, programs might be developed which speak to longer-tenured customers providing benefits to credit card users as well as benefits to high transaction users. Meanwhile, cluster 3 programs could be developed for the type of customer who is more financially sophisticated in terms of both investments and loans.

This type of approach allows marketers to use the information in a two-pronged manner: to both target the best customers based on a defined metric, but also to use the other information to communicate meaningfully with them.

Profiling vs. Clustering

Yet, as discussed above, in addition to the clustering approach, another option might be to profile the top billers (top 50%) and compare them with the bottom billers (lower 50%). The difference with comparing profiling with clustering is that the profiling approach utilizes an objective function, which is whether the customer is in the top 50% of customer billers. Other customer characteristics are then analyzed to identify which are the key characteristics in differentiating a top customer biller from the bottom 50%. We call this approach "supervised learning" because the objective function of being in the top 50% or not determines (supervises) what are the key characteristics in differentiating the top 50% of billers from the bottom 50% of billers.

In the clustering approach, there is no one variable which determines the outcome of the other variables. The objective of clustering is to identify those characteristics which best assign customers into unique and distinct groups. The objective here is to optimize not a specific variable or metric but rather a scenario where variation between customer groups or clusters is maximized and variation within customer groups or clusters is minimized. The data or variables in this case are "unsupervised"; that is, they are not analyzed against a specific customer variable or metric.

What, then, is cluster analysis? Cluster analysis represents a type of analysis which is considered unsupervised learning. In other words, any learning or insight from this type of analysis is not being directed against a certain piece of information. As discussed earlier, response models are considered supervised learning since the analysis is directed against the objective of trying to optimize response. Retention models attempt to find key triggers or variables which are directed against the objective of predicting retention. With cluster analysis, the learning is unsupervised in the sense that the analysis is attempting to uncover trends or patterns from all data with the objective being to classify individuals into distinct homogenous groups. The key difference from supervised learning is that the classification in cluster analysis is not determined or based on any one specific field, but rather on all the data fields.

How does cluster analysis work? In practical terms, cluster analysis attempts to classify records, usually individuals within the customer relationship management (CRM) marketing world, into distinct groups where each group is homogenous or similar in terms of their characteristics. In other words,

minimal variation of characteristics or behavior exists between individuals within the groups, but variation of characteristics or behavior between the cluster groups of individuals is maximized. This implies that the characteristics of one given cluster are indeed very different from the characteristics comprising the other clusters (Fig. 11.3).

Cluster analysis does not consist of just one technique. In fact, there are many, and ongoing research in statistics continues to explore the validation and value of new approaches. Currently, two of the most common techniques are K-means clustering and hierarchical clustering.

With both of these approaches, it is imperative that the data be standardized. Without standardization, two conditions arise that can input results in a negative manner. The first condition is outliers in the data, which means that results can be skewed by extreme values of a given variable. Clusters are created based on variance of the data. But this variance is determined or based on means or averages of all the variables in the data. If some of these variable means or averages are distorted by outliers, it would make sense that these distorted variables could yield misleading results in any cluster solution.

The second condition that can impact cluster results is the magnitude and scale of variable values, and the fact that magnitude of variables' values will vary dramatically in any cluster solution. Two of the common variables that might be part of a cluster solution are income and age. Income can range from 0 to millions of dollars while age typically ranges from 0 to 100. Again, cluster analysis is sensitive to actual values, and it is evident that the average or mean, and, of course, the resulting variance values, would be very different for income and age. Yet these differences would be based on the magnitude or scale of the data, rather than how each field truly varies relative to each other. Obviously, there needs to be a way to look at how data varies in a relative manner, thereby allowing the comparison of variables in an "apples-to-apples" manner rather than an "apples-to-oranges"-type approach.

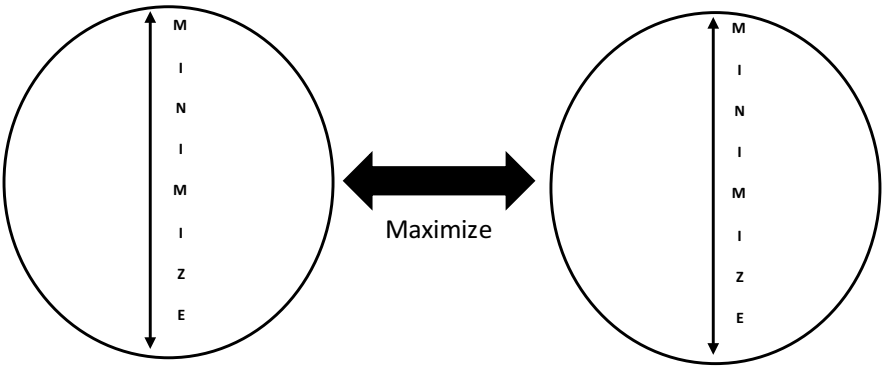


Fig. 11.3 Schematic of how cluster variation works

Both these situations, of outlier values and the magnitude of values, need to be adjusted prior to any cluster solution. The solution is to standardize the data prior to conducting any clustering exercise. There are several ways in which to do this.

Standardization can occur in a variety of ways. The most common approach is to normalize the data. This occurs by calculating the actual *Z* statistic variable value within the customer record. The actual numbers, in effect, range from −3 to +3, with larger absolute numbers being more statistically significant (+ or −1.96=95% confidence interval and + or −2.58 = 99% confidence interval). You will remember that the critical formula for calculating the *Z* statistic is:

$$\left(\text{Actual Value} - \text{Mean Value}\right) / \text{Standard Deviation}$$

This represents the foundation for most business applications in assessing the significance of a given business situation.

This formula essentially calculates the number of standard deviations. It is this standard deviation unit that can be compared across all variables. Variables which have tremendously differing scales, such as age and income, can now be compared or used in any analytical solution due to the standardization of values through this approach.

The second approach to normalization is less mathematical in the sense that each variable, with its range of values, is rescaled to a range between −1 and +1. Here's an example of what this means, looking at age (Fig. 11.4).

For age 25, algebraic calculation yields a normalized value of −.83 while for age 75, the normalized value is.75. This second approach is somewhat more limited in the sense that the distribution of records across this range of values is completely ignored. However, both approaches do provide options when it comes to normalizing the data with the preferred option being the use of the *Z* statistic. Most software routines today provide this type of capability.

Let's now look at the concept of clustering in more detail.

In clustering, centroids are produced which, in layman's terms, represent the center point of the cluster. The cluster and its centroid attempts to put records where the difference between the actual value of a given variable and the mean value of that variable are minimized within that cluster, while the mean values of that variable between clusters are maximized. The centroid consists of a multidimensional point which represents the means and averages of all variables within a given cluster. Remember that these cluster routines are iterative for a given cluster and its centroid. The iterative nature of the routine attempts to find observations with values that are closest to the centroid within

Age	18 19 20 21 100	
Rescaled Age	-1 -.99 -.98 ... +.97 +.98 +1.	

Fig. 11.4 Scaling age values to range between −1 and +1

its own cluster (homogeneity) while attempting to maximize the difference between the observation's values and the other cluster's centroid values (heterogeneity). In K-means clustering, the user defines the number of clusters whereby he or she analyzes the variance results of the defined cluster solution. The objective of the cluster solution is to explain as much of the variance of the data within the entire dataset.

But the question remains: what is this optimum number? This can be rather subjective given the underlying objectives of cluster analysis. If the perfect solution is desired, then the number of clusters would equal the number of records as 100% of all the variance is being explained by the solution. Obviously, no real insight or intelligence is provided here. There needs to be some sort of balance attempting to find the ideal number of clusters which provides meaningful insight and intelligence while attempting to achieve the cluster routine objectives of the maximal variation of information between clusters and the minimal variation of information within the cluster. One way to look at this is by exploring some notion of explained variation, such as the cubic clustering criterion or pseudo-F-statistic. As the number of clusters is increased, the % of explained variation will continue to increase until it is fully maximized when the number of clusters equals the number of records. As stated before, however, this is impractical as no insight or intelligence is provided with this so-called "perfect solution". The challenge then becomes to find precisely what this optimum balance is in terms of maximizing explained variation while providing the optimum level of insight and intelligence. One approach, the "elbow" approach, is to plot the level of explained variation across each cluster solution (Fig. 11.5).

As the number of clusters increases, the level of explained variation increases. From the below Fig. 11.5, each cluster solution is plotted against the level of explained variation. In terms of explaining variation, this curve begins to flatten out after a certain number of clusters. Identifying the point where the curve begins to flatten out is really the inflection point, which is a term with which mathematicians are familiar. It is at this point that one might identify the number of clusters as being the ideal solution. In the case above, this ideal number may be identified as being 4 since the % of explained variation begins to flatten

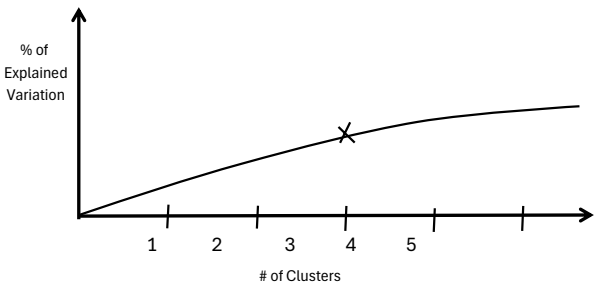


Fig. 11.5 Using the "elbow" technique to define the optimum number of clusters

out thereafter. In mathematical terms, it can be stated that the actual marginal increase in explained variation begins to decrease.

In the practical world, the analyst might focus on four clusters as being one option but other options, such as looking at a three-cluster or a five-cluster solution, would be explored. Again, however, these other options would be near the derived inflection point of the four-cluster solution. Once more, this is where the "art" of data science occurs. The analyst explores the means of all the variables within each cluster solution, as described earlier in this chapter. Judgment is then used by the analyst and the end business-user to arrive at a solution that makes the most practical sense from a business domain perspective, but one that would not deviate too far from the preferred scientific answer.

Rather than continue to delve more deeply into these techniques and its mathematics, it is more important to understand the overall mechanics of clustering and how it can create solutions for a given business initiative. Virtually all data science software packages contain clustering routines. In SAS, for example, one routine which quickly outputs a solution is appropriately named FASTCLUS. Here the user can quickly look at a variety of different cluster solutions where again the user is trying to find the inflection point at which the marginal increase in variance is maximized.

Once solutions are produced, the next challenge focuses on the communication aspect. In other words: what does this solution mean to business? Let us look at another example of a clustering exercise which was done for customers of a wealth management company (Fig. 11.6).

In this example, the final solution was a six-cluster solution, that is, a system that grouped their customers into six segments. We created descriptions for each customer segment by identifying those variables within the cluster that were statistically different from the other clusters. Keep in mind that we had more than 200 variables to consider in this exercise. To support our description, we reported the variable means of the statistically significant variables within the given cluster, and then compared them to the variable means of the other clusters. This was done in the case of all six clusters. In the Fig. 11.6, you are seeing just the result of the cluster 1 solution, and those variables which were statistically significant. You will observe that the mean of each of these variables differs greatly from the variable means in all the other clusters. This indicates that these variables can be considered as unique for this cluster. They can then be used to describe the characteristics of the cluster. In this example, the cluster 1 results provide the following profile:

Cluster 1 Variables that are significant								
Variable	Variable Description	Cluster 1 Mean	Cluster 2 Mean	Cluster 3 Mean	Cluster 4 Mean	Cluster 5 Mean	Cluster 6 Mean	Correlation Coefficient
MISSRISK	RISK TOLERANT	0.97	0.54	0.48	0.03	0.59	0.52	0.7303
MISSEMPLOY	OCCUPATION CODE IS MISSING	0.27	0.1	0.08	0.07	0.04	0	0.2735
LSTCHGYR	# OF YRS SINCE LAST CHANGE	4.11	2.86	2.25	2.83	3.81	3.16	0.2478
CGFpp	% IN CANADIAN GROWTH FUNDS	0.46	0.36	0.36	0.26	0.4	0.36	0.1968
REINVEST	REINVEST	0.02	0.01	0.01	0	0	0	0.0735
GGIFpp	% IN CANADIAN GROWTH INCOME FUNDS	0.02	0.01	0.01	0.02	0.02	0.01	0.0485
CONT2002	$ CONTRIBUTION IN 2002	58.06	355.23	319.1	203.54	203.54	135.64	-0.0621

Fig. 11.6 Wealth management company cluster 1 results

- Do not report their risk tolerance level (MISSRISK)
- Do not like to report their occupation on the KYC form (MISSEMPLOY)
- Have not had a change in a long time in any of their engagements with the company (LSTCHGYR)
- Put more money in Canadian growth funds (CGFPP)
- Like to reinvest (REINVEST)
- Put more in Canadian growth income funds with income (CGIFPP)
- Are less likely to contribute to an RRSP (CONT2002)

This same exercise is then conducted for all the clusters, with the end result being a profile description of what these records look like given that they reside in a certain cluster. In this cluster (cluster 1), the description might be that this segment is comprised of customers who are less engaged with the company, but more likely to reinvest, and doing this in growth funds but not in RRSPs.

The use of clustering can provide tremendous insight regarding the characteristics of a homogenous group of customers. Yet it is important to remember that business judgment is even more important in this exercise than in other data science exercises. This is because the description of the cluster is based on how a business user interprets results. There is a fair degree of subjectivity in relation to this. In a typical situation, the analyst and marketer will interpret the results and use their judgment to describe the cluster based on comparing the averages and/or median values across the different clusters. There is much more subjectivity in clustering-type exercises than there is in predictive modeling exercises. This is why cluster descriptions can be useful but, in many cases, do not reveal the complete picture of the cluster. In many cases, particularly in a multiple-cluster solution exercise, it will be discovered that descriptions used in one cluster will often overlap into descriptions of the other clusters. This reality provides a clear rationale for why one should look at cluster-type solutions as a very broad-based type of segmentation solution.

Key Takeaways

- Segmentation can be either simple or complex.
- Analyzing and explaining the variance of data is the underlying mathematics behind clustering.
- Besides the science of clustering, there is much "art" in arriving at the optimal number of clusters, yet practitioners use the "elbow" theory as a guide.
- Comparison of variable means between clusters is the approach used in describing the characteristics of a given cluster.

CHAPTER 12

Applying Data Science Techniques

The third stage represents the phase when the analytics (both advanced and non-advanced) occur. Segmentation, which was discussed in the previous chapter, would also occur in this phase. Predictive analytics' solutions, as well as business and campaign analysis, all occur within this phase. Arguably, this is where the "exciting" part of data science occurs. We have moved beyond the "data" phase which, in practical terms, is more critical than this phase in achieving project success. Data is the key ingredient, the "secret sauce", but now we begin to apply our tools to bring "magic" out of the data. We will commence this section by discussing several techniques.

Regression Analysis

One of the fascinating diagnostics that arises from regression analysis is output that relates to the power of a certain variable within an overall model. In the typical OLS (Ordinary Least Squares) regression technique, the diagnostic is often referred to as the partial R^2. Total R^2 represents the power of the model, with 1 representing a perfect model, where all the variability of the target variable is completely explained by all the input variables into the model. There is no random error. Similarly, a value of 0 implies that the model is completely ineffective in trying to explain the variation of the target variable. The variation of values around the target variable is all due to random error. Total R^2 essentially represents the explained variation accounted for by the model inputs divided by the total variation (explained variation + random variation). A partial R^2 represents the variable's contribution to the explained variance of the entire model. By dividing the partial R^2 into the total R^2, we can determine a diagnostic which is the variable's contribution to the overall model. These diagnostics can convey very interesting insights to marketers. Let's look at an example of a response rate model (Fig. 12.1).

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_12

Fig. 12.1 Contribution of variable to overall model

Model Variable	Impact on Response	Contribution to Overall Equation
Behavior Score	Positive	35.00%
Average Score	Positive	25.00%
Have an RRSP Product	Negative	15.00%
# of Fin. Inst. Products	Positive	10.00%
Avg. % of Credit Limit Used	Positive	10.00%
Live in Prairie Provinces	Negative	5.00%

Model Variable	Impact on Response	Contribution to Overall Equation
live in Quebec	positive	80.00%
tenure	negative	8.00%
Recency of last purchase	negative	7.00%
# of products purchased	positive	5.00%

Fig. 12.2 Example of a model's power explained predominantly by one variable

In this example, marketers are trying to sell premium credit cards to banking customers. In this case, the behavior score has the most impact (35%) in the model. Note how other variables related to engagement (i.e. banking activities) all exhibit a propensity towards buying a premium credit card. The notable exception to this is having an Registered Retirement Savings Product (RRSP), which may indicate a mindset that customers are "borrowers" or "seekers of credit" rather than "savers". Again, this is an opinion, but it does represent a notion that could be tested in a future campaign. The other variable (live in prairie provinces) indicates that customeers living in Prairie Provinces are less likely to purchase this premium credit card.

Now let's look at another example (Fig. 12.2).

In this example above, one variable (live in Quebec) accounts for 80% of the model's power. The variable itself is also binary, indicating that there are only two outcomes (1: live in Quebec or 0: do not live in Quebec). Under this scenario, we essentially have a one-variable model. But is this really optimal? A more reasonable next step would be to create two upfront segments:

- living in Quebec
- do not live in Quebec

The analyst would then develop separate models for each segment. In fact, this is one approach in how segmentation and modeling are integrated and, ultimately, how targeted solutions are best attained. But it was the use of R^2 statistics which provided the learning in determining these upfront segments.

In producing models, the ability to demonstrate the relative power and impact of each variable within a solution represents one feature that allows the marketer to "get inside" the "black box" of modeling. The notion of "black

box" means that it is not known how the key components of a given solution are determined. It is extremely important that not only data scientists, but also marketers and business analysts, understand the key components. Relying solely on the solution's ability to deliver results without some basic understanding of its components is a recipe for disaster. Furthermore, the ability to empower more people with the information on what comprises the model widens the knowledge base that is assessing the given solution. Based on this wide array of experience, the data scientist may find that a current model variable should not be used. For example, the purchase of Product X may be discovered as a model variable with a positive impact on response to Product Y. Upon further investigation, however, it may be discovered that for the last two years Product X has been given as a free incentive to new customers, and this incentive is now going to be eliminated. This information, provided by the business person or marketer, would suggest that this variable be eliminated in predicting response. As new customers seem to respond more, it would probably be very likely that a variable such as tenure would become a replacement variable or if it is in the current model, its overall power would be strengthened. These situations illustrate the point that this type of information should be shared with other business stakeholders in the project. Acknowledging that modeling is "black-box" technology with no understanding of the solution's key components is unacceptable in today's "knowledge is power" environment. Those organizations that can effectively communicate the technical details of a given solution to broader groups of stakeholders within the organization will develop solutions which are truly built by the team, rather than in isolation by the data science or marketing services area.

Factor Analysis

Factor analysis is a commonly used statistical technique for data reduction. Business applications include situations where the analyst may have hundreds of variables. Using factor analysis, the analyst can reduce the number of variables by using factor output scores as potential variables. For instance, one project may involve 200 variables, which are now accounted for by 17 factors based on the eigenvalue cutoff (the eigenvalue definition will be discussed later in this chapter). These 17 factor scores can be then used as input variables for an analysis or modeling exercise. Because of the difficulty in socializing these results, or in explaining the output in meaningful terms to the business stakeholders, the option of using factor scores as model variables is simply not used. However, the more common use of factor analysis is to select variables with the highest factor loading scores within a given factor. Figure 12.3 offers an example of this.

As you can see from this simplistic example, there are three factors which best explain the nine variables. In this example, the highlighted variables (income, education, and wealth) in factor 1 are the selected variables to represent this factor. In factor 1, one might easily describe the factor as representing

Variable	Factor 1	Factor 2	Factor 3
1) Income	0.905	0.255	0.255
2) Education	0.855	0.373	0.212
3) Wealth	0.956	0.303	0.185
4) Product A	0.303	0.855	0.205
5) Product B	0.295	0.805	0.245
6) Product C	0.323	0.755	0.285
7) Product D	0.105	0.355	0.755
8) Product E	0.155	0.405	0.705
9) Product F	0.085	0.304	0.725

Fig. 12.3 Factor analysis output

a measure of affluence. In the second factor, three products (A, B, and C) have the strongest factor loading coefficients. Upon further investigation, it may be discovered that these three variables essentially comprise a common product category of men's shoes. With this knowledge, a binary (yes/no) variable of men's shoes would be created as the variable which best represents the second factor. The third factor reveals that, once again, three products (D, E, and F) have the largest factor loading coefficients. In the third factor, these factor coefficients represent the product category of sporting goods. Using this knowledge, another binary variable (yes/no) would be created comprising the product category of sporting goods.

Factor analysis can prove extremely useful as an analytical tool. The user can vary the number of factors that are output by using different eigenvalue cut-offs. This flexibility allows the user to examine what kind of information and insights obtained with few factors versus many factors. As the number of factors increase, variables and information are observed, which are the most important in explaining a factor repeat themselves since they have already been used in explaining a prior factor. In examining the different factor number options, the user can then ascertain how much additional information is gained when adding factors and, conversely, how much information is lost when reducing the number of factors. As we have seen with much of the output and results from statistical analysis, the analyst uses both statistics and their own insights in deriving the optimal answer. The optimal answer is not totally driven by statistics, and nor should it be.

In defining the number of factors, the user utilizes the eigenvalue as a key statistical threshold in determining the number of factors. In most statistical software programs, this threshold is 1. The eigenvalue is a measure of variation where higher values reflect higher amounts of variation with the opposite happening for lower values. Accordingly if the value of 0 is substituted for the eigenvalue, the number of factors equals the number of variables input into the factor analysis routine as there is no data reduction since the explained variation threshold is 0. This is expected since an eigenvalue of 0 would imply that the routine is not utilizing any mathematics or statistics to perform any data reduction. As stated earlier, in any data reduction routine, although better insights and information regarding the analysis is gained, information is lost.

The eigenvalues enable the user to determine how much total information is being lost through the use of fewer factors. As stated before, the eigenvalue is a statistical diagnostic which captures the total information which is captured by the given factor through its ability to explain all the data in the sample. As you add in more factors, each subsequent factor has a smaller eigenvalue, thereby indicating that it is capturing a smaller % of the explained variation of all the data in the data set. One could plot the total eigenvalue against each factor number option or principal component, as seen in the chart below.

Similar to how we determined the optimum number of clusters, we are trying to find the inflection point, where the increase in loss of information becomes marginal, and use this as the optimum number of factors. It is at that point that adding in incremental factors does not produce a noticeable decrease in the eigenvalue. In other words, the marginal impact of adding more factors is not really providing real information gain. From the chart depicted in Fig. 12.4, this could be principal component or factor 10. Once again, the raw statistical output will be used in conjunction with the analyst's own insights and experiences which are drawn from fewer factors versus more factors. These insights and experiences may, in fact, result in the analyst selecting the optimum number to be either less or more than what science might suggest. But the analyst's selected number of factors will not deviate too much from the optimum number of factors determined from the graph above because the optimal data science answer is a combination of both science and practical insights. In most situations, the analyst will collaborate with the business person concerning his/her insights from the results. This type of process is similar to that which was discussed during cluster segmentation.

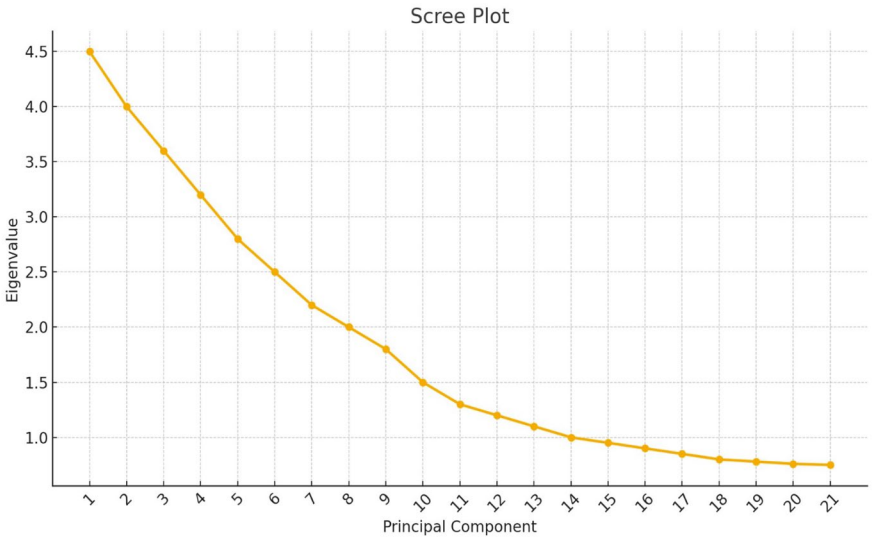


Fig. 12.4 Example of scree plot in determining optimum number of factors

Users can apply these techniques in predictive modeling scenarios where there is a significant opportunity to reduce the number of variables. The same scenario would also apply to clustering, where the user may be confronted with hundreds of input variables. In fact, the use of factor analysis is a standard mandatory procedure within clustering since clustering routines become less robust as more variables are introduced.

Routines like factor analysis can be used to determine the reduced set of variables (40–50), which is the typical range of variable inputs into any clustering exercise. Within predictive modeling, factor analysis is an option and not a requirement in building the solution. The most commonly used technique in factor analysis is called principal component analysis (PCA). There are other approaches to factor analysis, yet they all have the common objective of reducing data and achieving this through what is commonly referred to as unsupervised learning which has been discussed earlier. In other words, the technique maximizes the variation of the data set into patterns, and not against an observed behavior. By not maximizing this variation against an observed behavior like a predictive response model, it is said that the learning (or derivation) of the factors is unsupervised. Another important fact to remember is that each pattern or factor is completely independent from each other. No multicollinearity (i.e. no correlation) exists between factors. In determining the number of factors that are actually produced from the routine or algorithm, the user has control of this through the values of the statistical criteria that he or she inputs into the routine, such as the eigenvalue metric that has just been discussed over the last few pages.

As in most statistical routines, debate surrounding these routines continues amongst academics, many of whom maintain that factor analysis is limited by its assumption of linearity. Most distributions of data are not purely linear. Ongoing research is attempting to produce data reduction routines that are nonlinear in nature. From a pure mathematical or academic purist standpoint, nonlinear routines will be superior to the linear ones. From a business standpoint, however, the real advantage will be the incremental value that this provides to a given data science problem compared with what is produced by the traditional factor analysis conducted on the same data science problem. As always, this incremental advantage in the business world will be measured by the additional $ benefits accrued to the new technique.

Developing the Solution

Previous chapters have discussed the very critical role of creating information that can be used in any data science solution. The derivation of new fields and variables represents a key area of the data scientist's insight and expertise within the overall project. In fact, as stated in previous chapters, the creation of this data environment represents by far the most time-consuming part of the exercise. This makes sense as it is within this stage that the data scientist will fundamentally determine the quality of a given solution. The use of a variety of

different statistical techniques will, in many cases, yield the same performance. Yet, in most practical situations, it is the data, rather than not the mathematics or the statistics, that determines the performance of a given solution.

Recognizing the importance of data, and the ultimate creation of the analytical file, there is still the need to do something with the data. In other words: how does the data scientist want to work the data? Do they want to produce reports or conduct some ad hoc analysis or engage in more rigorous forms of mathematics in terms of developing more sophisticated solutions, such as models, profiles or cluster segments?

Before conducting any analysis or developing more sophisticated solutions, the goal of the data science project must be clarified. For instance, is the analysis trying to gain insights and learning against a given customer behavior? If so, the analysis is being directed or supervised against this behavior; this is often referred to as supervised data science. In cases where the analysis is directed or supervised against a given behavior, data scientists will often refer to this as the "dependent variable" or "objective function". Other situations require analytics, but the analysis is not directed against a given objective function or dependent variable. Examples of this include cluster analysis and factor analysis, which have already been discussed. Suffice it to say, both of these technologies (cluster and factor analysis) can analyze reams of data (often hundreds of variables at a time); in both cases, however, these analyses are not being directed by any specific customer behavior. In more practical terms, if the data scientist is requested to conduct a customer attrition analysis, then the data science exercise is seen as directed, as they will identify a specific field that relates to attrition such as "Cancel Reason". They could then potentially analyze all other customer-related fields against the above-mentioned field. A similar approach is also adopted when looking at all other types of predictive models. These types of model require that the user identify what the model is trying to predict. For example, a predictive response model would require that the analyst create a field on the analytical file that indicates response (1: yes, 0: no). The response field in this case becomes the dependent variable or objective function of response. In simpler terms, all the analyses in building the model are directed at trying to find information which best predicts the objective function of response.

Having been clear about what type of analysis is required, the data scientist is now in a position to begin using the variety of tools that are available within this stage of the analysis. The choice (or selection) of a certain tool will depend on the specific business need. For instance, in many cases, the analyst is simply asked to produce reports. No sophisticated statistical algorithms are required as the technical need is simply to manipulate and organize the information in order to produce the required reports. In building a predictive model, however, a variety of statistical tools might be used, and the selection of a given tool will depend on the specific task that the data scientist is trying to perform. For example, if the exercise is a model-building one, then the first requirement might be to obtain some sense of what variables and information most impact the desired modeled behavior.

	Live in Quebec	Spend	Number of products	Recency of Spend	Credit Score
Live in Quebec	1	-0.5	-0.65	0.3	0.4
Spend	-0.5	1	0.7	-0.45	0.5
Number of products	-0.65	0.7	1	-0.52	0.55
Recency of Spend	0.3	-0.45	-0.52	1	-0.24
Credit Score	0.4	0.5	0.55	-0.24	1

Fig. 12.5 5 × 5 correlation matrix

The use of correlation routines serves as the first statistical tool that can yield these initial insights. These correlation reports, though, do not represent the typical correlation matrix reports with which most statisticians are familiar. Traditionally, if one is conducting a correlation analysis on five variables, one can expect a 5 × 5 matrix (Fig. 12.5).

In this correlation matrix, we observe the relationship between all possible variable pairs. The strength of the relationship is determined by the absolute value of the coefficient correlation, as depicted in each cell of the matrix. The sign of the coefficient indicates simply whether the trend between the two variables is negative or positive. In our above example, the strongest relationship occurs between spending and the number of products, with a correlation coefficient of .7. As spending increases, the number of products increases. Meanwhile, the strongest negative relationship (−.65) was between having an RRSP and number of products, indicating that people that had an RRSP had fewer overall products. Note that correlation coefficients of 1 in this above matrix were depicted for those variables being correlated against themselves.

Yet, in data science, we do utilize the concept of correlation analysis. Rather than exhibiting a complete matrix as seen previously, however, we design the report where we analyze all variables against the dependent variable or objective function.

For example, if we had seven variables that included response behavior which is the behavior we are trying to optimize, then the correlation report would consist of a 6 × 1 report. Here the variable to be predicted (response) is listed on the horizontal axis and all the other variables (independent) are listed on the *y*-axis. In this report, all variables on the *y*-axis are ranked by the absolute value of the correlation coefficient, against the variable listed horizontally on the *x*-axis. As stated above, negative values are equally as good as positive values. Besides the correlation coefficient, which is used to rank the other variables against the variable listed horizontally, we also observe another column, called the confidence interval. This confidence interval is a metric which relates to the statistical significance of the variables on the *y*-axis versus the variable on the *x*-axis, or our dependent variable. Basically, the larger the correlation coefficient, the larger the confidence interval. The sign of the variable indicates whether there is a positive or a negative impact with the *y*-axis or dependent variable. Figure 12.6 is a correlation matrix used to develop a predictive response model. The *y*-axis variable is response and does not need to be shown

Variables	Correlation Coefficient	Confidence Interval
Age	-0.0673	0.995
Tenure	0.055	0.98
# of products purchased	0.045	0.97
# of promotions since last purchase	-0.031	0.96
Income	-0.0045	0.5
Household Size	0.001	0.2

Fig. 12.6 Correlation matrix of six variables versus response

since the table implies that the correlation results are being measured against response.

In this simplified example above, the report indicates that age, tenure, # of products purchased, and number of promotions since last purchase are statistically significant against response. Meanwhile, household size and income are not statistically significant against response and would not be relevant in any modeling solution. In this example, the analyst is attempting to build a response model. The results of the correlation analysis would indicate that the statistically significant variables should be considered in any data science solution. Clearly, any data science solution would certainly explore the notion that younger people with longer tenure, and who have bought multiple products but who have not been promoted most recently, are more likely to respond. With correlation analysis, the analyst essentially has their initial mathematical glimpse of the data. Mathematics begins to provide quantitative support in formulating any data science solution. However, the analysis is far from complete at this stage. The further manipulation of variables can occur based on the insights of the correlation analysis. Suppose that one variable, such as age, had a correlation coefficient of .3 while all the remaining statistically significant variables had correlation coefficient values of .10 or less. Due to the scale of difference between age and all the other variables, this information might tell us that segmentation based on age should occur prior to the development of any models. CHAID (Chi-Square Activation Interaction Detection), which will be explained more fully in later chapters, is often used as a decision-tree tool in building models. As the tool can be used to define the actual variable breaks within a branch level, it can also be used to define the optimum breaks for a given variable. In our example above, CHAID would be used to determine the optimum age breaks which would then be used as our segment definitions (Fig. 12.7).

For each segment, separate models could be developed. Yet this insight and the use of CHAID might never have been considered if the correlation analysis had not revealed the age results.

Beyond correlation, one might consider the use of CHAID as another tool in deriving new variables. What is meant by this? As previously mentioned,

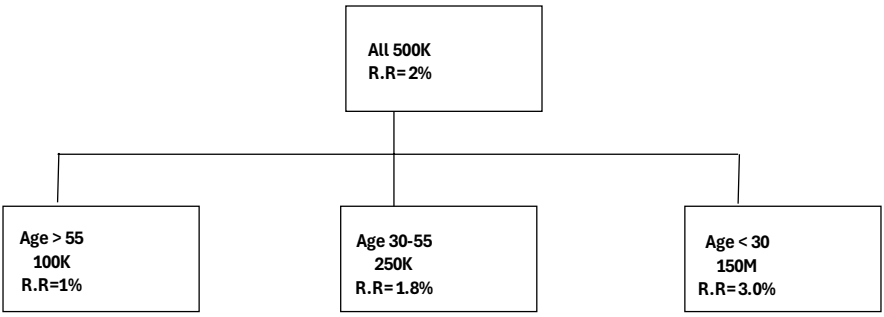


Fig. 12.7 Example of CHAID used to define model segments

variables with many different character value outcomes (otherwise referred to as data granularity in the negative sense) are not useful in any data science exercise. One example is product-level information, where there could be hundreds of different products that are sold. In a typical data science manner, hundreds of binary variables would be created (1 if a product was purchased; 0 if a product was not purchased). As discussed earlier, many of these newly created binary variables would be meaningless in any data science exercise. The number of occurrences within that variable (the 1's in this case representing customers who have purchased that specific product) would be very sparse and very likely insignificant in any data science exercise. Let's look at what this exactly means.

As a first step, the analyst would typically create binary variables for each potential outcome. If there were 200 product categories, 200 binary product variables with "yes/no" (1/0) outcomes would be created. Now suppose we had 200,000 customers. Calculating an average product purchase using just the above information, we might estimate .5% of the customers, which represents that, on average, 1000 have actually purchased a product from the list of 200,000. This .5% is simply too small to identify any meaningful trend or pattern.

But what if there was a way to group character values together? Suppose, in the case of a retailer, products could be grouped together into some broader category. By grouping these occurrences together, in essence there are more 1's (occurrences of purchase within the broader product category) or critical mass around a given occurrence.

The use of CHAID provides the key tool in being able to group these outcomes in an optimum manner. The grouping of these outcomes would be based on the behavior which is attempting to be optimized. In the example above, it would be response.

Figure 12.8 is an actual CHAID example of product categories.

For the purposes of simplicity, only 10 product codes are used in this example. These are represented by the letters A–J. In this example, we observe how product codes (which are in letters) are grouped into optimal broader

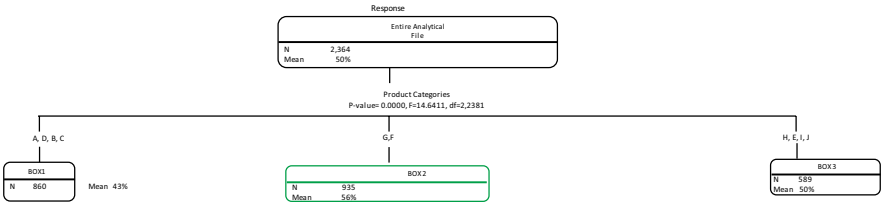


Fig. 12.8 CHAID used to group product codes into categories

categories when analyzed against response. With this information, binary variables (0, 1) could be created for each node. A superior approach, though, would be to create a product index variable with three outcomes:

- 1: For the node representing G and F product categories
- 2: For the node representing H, E, I, and J product categories
- 3: For the node representing A, D, B and C product categories

This ordinality (1,2,3 outcome values) is determined simply by ranking the nodes based on the response rate mean within each node. By ordinality, the lowest response rate node (Box 1: A, D, B, and C) is assigned a value of 1, the next highest response rate node (Box 3: H, E, I, and J) is assigned a value of 2, while the remaining node with the highest response rate (Box 2: F and G) is assigned a value of 3.

As we have observed in this simplistic example above, new and more meaningful variables can be derived in situations where character variables have many different outcomes. This is very prevalent within the retail industry, where it is not unusual to observe over a hundred different product codes (or SKUs) within a given institution.

The use of CHAID and factor analysis demonstrates the notion that variable creation can continue even once the analytical file has supposedly already been created. Variable creation will continue if there are insights into how best to manipulate information. New information and variables are easily added onto the analytical file based on the learning arising from stage 3 of the data science process. But what happens next within the process?

Along with using the above-mentioned statistical routines, the discovery process within this stage uses routines to allow the data scientist to better understand the information that might comprise a given solution. This is certainly even more necessary if the data scientist needs to communicate the results to a business end-user. Examples of such routines are called EDAs (Exploratory Data Analysis) reports. An example of such a report is given in Fig. 12.9.

In this first example of the above table, which is an attempt to build a response model, the variable, tenure, has a strong positive relationship with

Tenure in Years	% of Customers	Response Rate	Response Index
0-1	25.0%	2.0%	57
2-4	25.0%	3.0%	86
5-7	25.0%	4.0%	114
7+	25.0%	5.0%	143
Average	100.0%	3.5%	100
Age	**% of Customers**	**Response Rate**	**Response Index**
Under 25	25.0%	6.0%	171
25-35	25.0%	4.5%	128
35-50	25.0%	2.5%	71
50+	25.0%	1.0%	28
Average	100.0%	3.5%	100
Household Size	**% of Customers**	**Response Rate**	**Response Index**
1	25.0%	4.0%	114
2	25.0%	3.0%	86
3	25.0%	4.0%	114
4+	25.0%	3.0%	86
Average	100.0%	3.5%	100

Fig. 12.9 Examples of exploratory data analysis (EDA) reports for tenure, age, and household size

response. In other words, the higher the tenure, the more likely the person is to respond.

In this second example of the above table, older people are less likely to respond, indicating the negative relationship between response and age.

In the third example of the above table, the report indicates that there is no relationship between household size and response.

With several diagnostic tools already being used here to develop models, such as correlation analysis, factor analysis, CHAID, and EDA reports, the analysis is now ready to employ more robust statistical routines that will, in fact, create the final model. The analyst now understands the information that is being input into the statistical techniques. Analysts must understand the information that is being considered for a given model. Just simply throwing raw data and information into a model-building technique is a recipe for disaster. Given this deep understanding of information which can potentially be used as inputs into the model, a number of different techniques can be employed. There are no shortages of mathematical techniques for the analyst. Certainly, however, within the marketing arena, traditional techniques such as logistic regression and linear regression work quite well in most cases. Some of the other and more advanced techniques, such as neural nets, genetic algorithms, etc., are other options. The difficulty with using these options is that they are very complex and difficult to understand, particularly with regards to the output. This is not to say that they should not be considered. After all, if these options can demonstrate clearly superior results (performance lift) to the

more traditional type of techniques, then they would be appropriate. Again, however, data science experience has demonstrated that many of these advanced techniques simply do not significantly outperform the more traditional techniques. The reason is that these more advanced techniques require tremendous volumes of data and, more importantly, the data needs to be less noisy or have smaller portions of unexplained variation of the data. This is an important concept as in other business applications, such as optimizing the degree of quality control within a manufacturing process, we observe much less noise or "unexplained" variance The unexplained variation is indeed quite small, which implies that much of the data can be explained. Ultimately, these more advanced-type mathematical techniques work well when compared with the explained data in trying to achieve the best solution. These concepts apply extensively in our world of GENAI models that are being built extensively with huge computers and massive volumes of data. The issue of "noise" becomes mitigated since there is less randomness with text and image data. So, before I discuss what should happen next, it is important to remember that these models (advanced neural net or deep learning models) should, and must, be used in these situations. In developing NLP models for GENAI solutions, this type of mathematics represents the best and only option.

The world of consumer behavior is different, however. In the marketing world, large random unexplained variation is a fact of life when trying to explain consumer behavior. Trying to explain away this random variation can lead to results that appear to be good during the development of the solution, but which, when applied within the real world of business, lead to performance that is always sub-optimal. This situation in data science is often referred to as an overstatement of results. These techniques must be used judiciously, meaning that the use of validation samples is even more imperative. Otherwise, results achieved within a marketing campaign will always be inferior to what was achieved during development.

The more traditional techniques, because of their less complex nature, simply don't have the capability to overstate results. This can be easily demonstrated by the typically low F values and R^2 values that are observed once a final model is created. Much of the actual variation within a final model is indeed unexplained. Obtaining an R^2 value of .05 might represent an optimum model-building scenario for any data scientist given the current data environment. But remember our threshold of success is not statistical diagnostics, but rather gain charts/decile charts or Area under the Curve (AUC) curves, which are discussed later in this book.

Given that these more traditional techniques are being used, what happens, and how are the results communicated to end-users? The user will now conduct logistic or linear regression on the analytical file whereby the user defines the objective function or the behavior that is attempting to be optimized and the independent variables as those variables which are deemed to be potential model predictors of the desired behavior. A series of regression analyses are run against the data, whereby each regression equation represents one option given

a certain set of variables. The objective in running these series of regression routines is to eliminate variables such that one final model is provided in which the best 10–15 variables are run as the last regression equation.

Typically, it is not unusual to run 5–10 of these regression analyses prior to the development of a final solution. The number of these regression analyses will, of course, be determined by the learning and insights from correlation analysis, CHAID, and EDAs. The first tool used to prescreen the variables, as previously discussed, is correlation analysis, where the analyst will use a statistical significance level threshold of 95% to determine which variables should still be considered in developing the final model or solution.

It is this group of statistically significant variables from the correlation analysis, plus other insights from both CHAID and the EDA reports, which can be used to identify variables that can be used as inputs into a regression routine (iteration). However, the analyst still needs to find a way to effectively analyze all these variables in the development of the model. Using correlation analysis, the analyst can formulate groupings (iterations) of variables based on the relative statistical strength of each variable. In determining what variables to include in all the groups, the decision is more scientific as the analyst decides to exclude variables that are below a certain statistical confidence interval based on results from the correlation analysis. This subjectivity occurs when the analyst determines how many variables there should be in each grouping. One rule of thumb is that no more than 15 variables should be included in any regression equation. By including all variables (such as 70+ variables, which is not an unlikely situation) based on EDA and correlation results, the argument is that information is lost by performing only one routine against all 70 variables. With 70 variables in one routine, some of the lower-impact variables may be excluded from the model. This may occur because variables with low correlation, but which are still statistically significant would not be in the final model. The analyst accordingly will group together variables (ideally 15 per group) based on their strength. In the initial phases, variables of comparable strength (correlation coefficient values) will be analyzed against each other within one group. In effect, a variable with a very low correlation coefficient, but which is still statistically significant, might end up in the final model because it has low multicollinearity with the other variables. This effect of running a series of stepwise regressions mitigates the impact of any lower-impact variable being excluded from a final model, at least during the initial run of regressions (Fig. 12.10).

Once again using the output of correlation, and CHAID, we have determined that the 40 remaining relevant variables from our preliminary analysis should be gathered into four groups. These four groups are ranked in terms of correlation strength, with group A being the strongest and group D being the weakest. Some “art” can also be used to eliminate variables that are known to be correlated, thereby extracting variables that have a certain level of uniqueness or “independence” from each other. In our example here, if one regression routine is performed on all 40 variables, it is highly unlikely that any

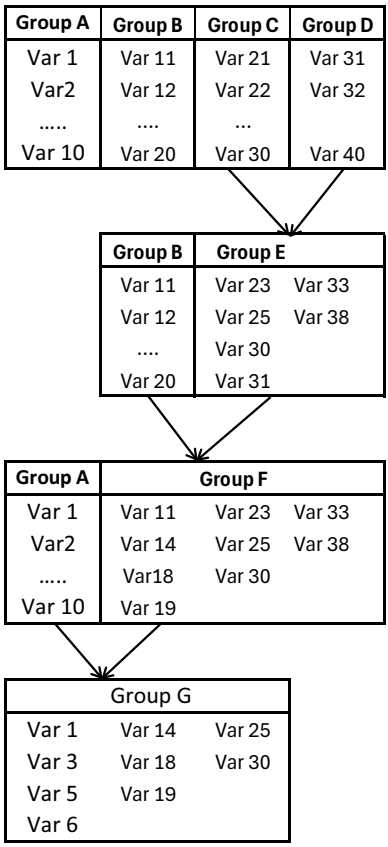


Fig. 12.10 Example of stepwise iterations to develop final model variables

variables from Group D will make it into the final model. They are excluded because of the overwhelming influence (i.e. statistical strength) of Group A variables. By performing a series of regressions (i.e. one regression for each above group), the number of variables within each group is reduced based on their interaction with each other, as well as their impact on the objective function or dependent variable.

In our example above, the first regression is on the weakest variable groupings (Group C and Group D). The surviving variables of Group C and Group D, called Group E, are then regressed with the variables of Group B. The surviving variables, Group F, are then regressed with the strongest group (Group A), with the final model variables being Group G. Notice how the approach here is to analyze the weakest variables first in order to identify survivors and to then see how they perform against the strongest variables.

Through this approach, it is more likely that variables from the initial Group D can actually make it into the final model. Since certain variables are excluded primarily as a result of multicollinearity, especially with the higher groups (A and B), there are fewer variables from these stronger groups (A and B) that can

compete with the lower group. In some cases, variables from the lower group (Var. 25 and Var. 30) can bring in certain information into a model which indeed impacts the overall solution and is somewhat unique relative to the other model variables (low multicollinearity).

With this type of approach of using a series of regressions, a final model is eventually produced. But what does this look like, and how are the results communicated to business end-users? The approach used here results in the development of a solution which is parametric. But why is the notion of parametric vs. non-parametric important to understand? Because parametric implies that each variable has a weight or coefficient assigned to it which is determined from its impact with the modeled behavior, as well as its interaction with the other model variables (multicollinearity). In the case of non-parametric solutions, there are no assigned weights or coefficients to each model variable. The variables themselves create the solution; they are not influenced at all by multicollinearity, which implies that there is no need for weights or coefficients to be assigned to variables. In effect, these types of non-parametric solutions represent a series of business rules. A good example of such a solution is CHAID, which has already been discussed, whereby the solution represents those optimal segments which have been determined by a series of business rules.

Meanwhile, the parametric solution is comprised of an equation with weights such as:

$$\text{RESPONSE} = .008 + .0015^{*}\ \text{Income} - .01^{*}\ \text{Age}.$$

Age and income are parameters, but the solution is parametric due to weights or coefficients assigned to each parameter (+.0015 to income and −.01 to age). The actual weight or coefficient is determined based on both the impact of the variable with the objective function such as response, its interaction with other variables (multicollinearity) and the magnitude or scale of the variable. For example, age and income will have two different coefficient values. In a response model, age, which can vary reasonably from 0 to 100, will have a much larger coefficient than income, which might range from 0 to millions of dollars. Yet, the coefficient will also increase or decrease in absolute terms based on its impact with response, where increase in absolute value implies a stronger significance with response, with the reverse holding true for a decrease in absolute value. Additionally, strong multicollinearity with other variables will also reduce the parameter or coefficient of a given variable.

With an actual equation produced, however, what is presented to business users or management? Certainly, the equation itself might be of interest to the more technical or advanced data science user. To other people within the organization who have a vested interest in the model, however, communication of the results needs to have more of a business focus. This business-oriented communication needs to focus on two areas:

- Description of the solution
- Impact to the business

Figure 12.11 represents a model with six variables or parameters ().

The table lists each variable in order of priority or significance with the model. This can be seen in the column "Contribution to Overall Equation", where the first variable (Behavior Score) accounts for 35% of the model's overall power and the sixth variable (Are Female) accounts for just 5% of the model's power. The column called "Impact on Response" describes the relationship or trend of a given variable with response. For instance, the most important variable (Behavior Score), as attested by its 35% contribution to the overall model equation's power, indicates that persons with higher behavior scores are more likely to respond. Conversely, the third-strongest variable (15% contribution to overall equation) indicates that persons with an RRSP product are less likely to respond.

With this kind of information, the general characteristics of the model solution can be described in business terms. For instance, the information in Fig. 12.11 would indicate that responders have the following characteristics:

- Have higher behavior scores
- Are higher spenders
- Are not likely to have an RRSP
- Have a variety of financial products
- Are heavy credit revolvers
- Are not female

Now, the same kind of approach is also used in other strategies.

The key to success in using the variety of statistical techniques that are now commercially available is the ability to extract the relevant information that will yield meaningful business insights. In regression modeling, it is important to know what the key characteristics of the model are, and what their relationship is to the objective function. Clustering solutions yielded insights that allow business users to understand the key drivers or characteristics of a given cluster segment, whereas factor analysis provides a means to summarize or reduce our data into a smaller, more manageable set of variables for analysis.

Fig. 12.11 Example of final model variable report

Model Variable	Impact on Response	Contribution to Overall Equation
Behavior Score	Positive	35%
Average Spend	Positive	25%
Have an RRSP Product	Negative	15%
# of Financial Inst. Products	Positive	10%
Avg. % of Credit Limit Used	Positive	10%
Are Female	Negative	5%

Data science is not only about extracting trends; it is also about extracting meaningful business trends that can be actioned within some business initiative. The use of statistics enables data scientists to deploy some science to the data to further reduce the degree of subjectivity within an analysis, recognizing that subjectivity within a business environment will always play some role within the overall solution. In fact, in most cases this subjectivity plays a significant role as it is founded on the domain or business knowledge expertise of the various key stakeholders within the given organization. In many cases, the "science" confirms our hypotheses gained from domain knowledge.

Key Takeaways

- Once the analytical file, several statistical analyses can be done prior to the use of any specific modeling techniques.
- Techniques like factor analysis are unsupervised, but they do help the practitioner by reducing the data or the variables that need to be considered in a modeling exercise.
- Techniques like CHAID are supervised, but help the practitioner to create even more variables in our analytical file.
- Examination of EDA reports, which is supervised, can also yield insights on the creation of new variables in our analytical file
- Techniques like correlation analysis are supervised, but help the practitioner in filtering those variables which should be considered in the modeling analysis.
- A number of modeling techniques should be considered; in the world of GENAI, deep learning models are the only option.
- In trying to predict consumer behavior, more traditional and simpler models work quite well and are easier to explain.
- In using regression models, one approach has been to utilize a series of stepwise routines in producing a final model with the end deliverable being a final model variable report that explains the given model solution.

CHAPTER 13

Gains Charts and AUC Curves

Many different solutions are produced in data science. As previously discussed, solutions can be either statistical or non-statistical. But how is the impact of a given solution to be assessed? In particular, how is this impact assessed in business terms which are readily understandable? The answer is through validation by what is often referred to as a gains chart or Area Under the Curve (AUC) curve. Yet, it should be noted that these types of validations are only appropriate for exercises focused on consumer behavior. In building models using image recognition or natural language processing (NLP), accuracy and loss rates are the only appropriate validation measures. These will be discussed in more detail in the chapters dealing with deep learning and neural nets (Fig. 13.1).

Gains chart tables actually work in a very simple manner. If a solution is being produced against a given set of observations (such as, for example, customer records), the solution is applied to this set of observations. The set of observations can then be ranked by the solution into groups which are most often either in the form of deciles (10% intervals) or half-deciles (5% intervals). The top group or interval would represent the group identified as being the highest priority in terms of the solution while the lowest group or interval would represent the group with the lowest priority. Against these intervals, the observed or actual reported metric is determined which is being optimized within a data science project. For instance, in the response model above, the cumulative average response rate for each interval (10% deciles, for example) would be reported. The next column would report the % of all responders within the entire sample that is being captured within the interval. Other columns that would be included are metrics such as the cumulative response rate index, which essentially comprises the lift or the ratio of each interval's cumulative response rate relative to the average response rate for the entire sample. Alongside these columns, one could also include the return on investment (ROI), if cost per promotion effort and revenue per sale are available. The final

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_13

% of Validation Sample	Validation Names	Cumulative Response Rate	Cumulative % of Total Responders	Response Rate Lift	Interval ROI	Modelling Benefits
0-10%	20,000	3.50%	23%	233	145%	$26,667
10-20%	40,000	3.00%	40%	200	75%	$40,000
20-30%	60,000	2.75%	55%	183	58%	$50,000
30-40%	80,000	2.50%	67%	167	22%	$53,333
40-50%	100,000	2.25%	75%	150	-13%	$50,000
.	.					
.	.					
.	.					
90-100%	200,000	1.50%	100%	100	-58%	$0

Fig. 13.1 Gains chart example

Fig. 13.2 Calculating the modeling benefits at 0–10% of the gains chart

	# of responders	Promotion Costs
With Modelling	700	$20,000
Without Modelling	700	$46,667
Modelling Benefits		$26,667

column is what we refer to as the $ opportunity cost, which represents the dollars saved because of the data science solution. Let's look at how this is exactly calculated by considering the first row (0–10%). By determining the number of responders that were achieved with modeling in this first row ($700 = .035 \times 20{,}000$), we now want to determine the incremental marketing dollars which would have been incurred to achieve this same responder quantity. Now, however, we do so without modeling. Figure 13.2 gives a schematic of how this would be calculated assuming a promotion cost of $1.00 per effort.

Note that the $46,667 number is reached by simply dividing the overall response rate of the population (1.5%) by the number of responders that are captured by the model in the top 10% ($700/.015$), and then multiplying this result by $1.00.

In evaluating a model or solution through this gains table, analysts primarily look at how well the given solution rank orders the desired objective. In the case of a response model, it is optimal to observe a response rate which is rank-ordered very strongly from the top decile (0–10%) to the bottom decile (90%–100%). Another way to review the gains table is by plotting the interval response rate of each decile with deciles on the *x*-axis and interval response rate on the *y*-axis (Fig. 13.3).

This is referred to as a Lorenz curve, and the steeper the line, the better the solution. A flat line would indicate that the solution has been completely ineffective.

Another common view that many data scientists use when evaluating solutions is by plotting the cumulative % of the desired solution on the *y*-axis against the ranked intervals on the *x*-axis, with 1 being the top rank and 10 being the lower rank. This often referred to as the AUC curve (Fig. 13.4).

The area between the two lines (curve line and straight line) is the measure of performance. and see below sentence which talks to the measure of performance.

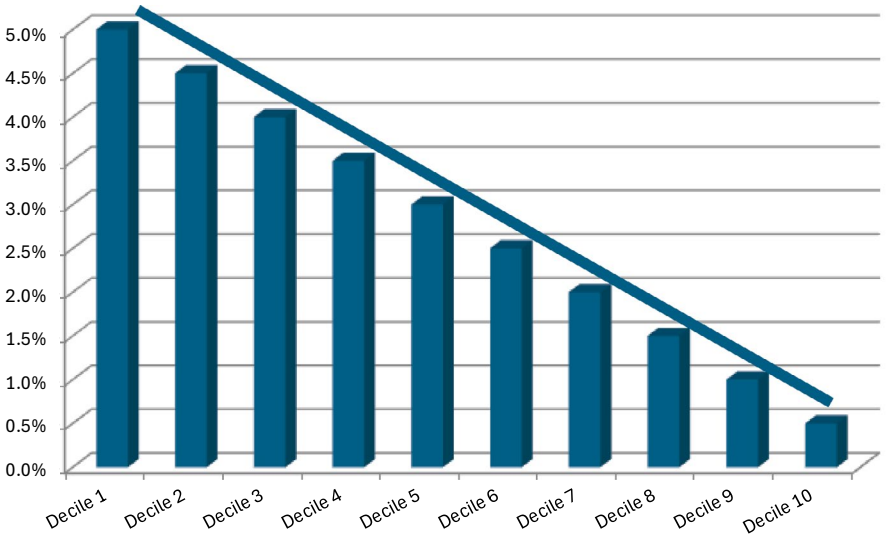


Fig. 13.3 Lorenz curve plotting interval response rate versus model decile

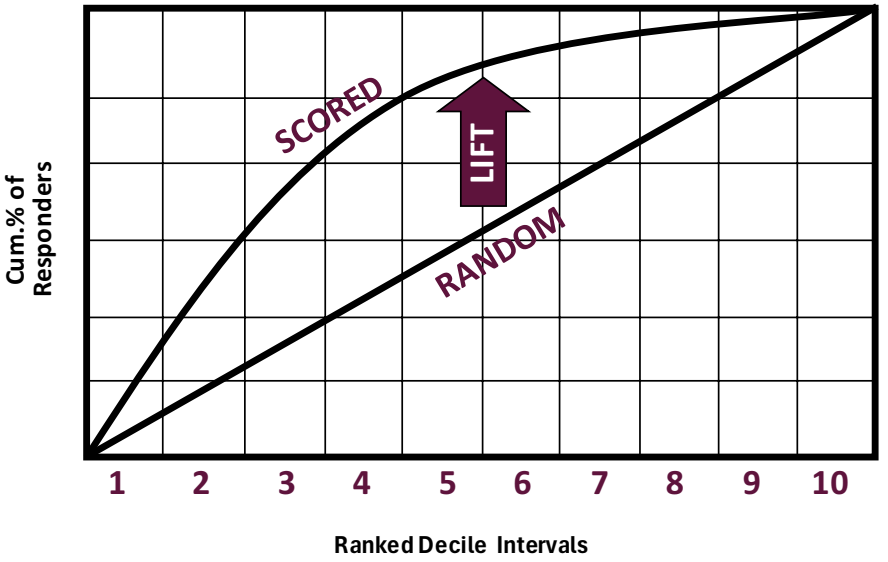


Fig. 13.4 AUC curve: Plotting the cumulative % of responders versus model decile

In attempting to place a value on a given solution, data scientists will often use this AUC curve depiction to measure the space between the straight line and the parabola. They see this value as a measure of a data science effectiveness. The two mathematical quantifications used in calculating this space are

the Kolmogorov-Smirnov statistic (more often referred to as the KS statistic) and the Gini coefficient. A high KS statistic indicates a better-performing model; a Gini coefficient close to 1 indicates a very strong model with values close to 0 reflecting a weaker model.

In evaluating models, many institutions will look at the degree of accuracy between the observed metric and predicted metric. In the case of response rate models, some statisticians will regard the difference between the predicted and actual response rate as the primary focus. Although a minimal difference between the predicted and observed usually produces a good model, it is not the primary criterion in assessing consumer behavior models. Rank ordering of the observed behavior, which in this example is the response rate, is the key performance criterion here. Often enough, various statistical techniques will be employed in a given solution. For example, it may be both linear and logistic regression are used in a particular solution. In the case of binary outcomes such as response, theorists will argue that logistic regression should be used as its expected outcome is binary. Yet, if rank ordering the behavior is the most desired performance metric, then using either linear regression or logistic regression is equally valid. However, if the predicted outcomes of the binary solution are required as part of a more comprehensive model, then the use of logistic regression is a necessity. One good example may be when we need the predicted probability to calculate some more comprehensive expected value behavior.

If rank ordering is the desired behavior, it may be observed that even less robust techniques such as CHI Square Activation Interaction Detection (CHAID) may deliver an equally acceptable solution. The actual output from CHAID represents a number of groups of customers in which each group can be ranked based on the observed behavior. This is less robust than techniques, such as linear regression or logistic regression, where the actual output is an individual score for each customer record. However, when comparing these less robust techniques such as CHAID, it is found that, in some cases, it can be shown that the desired objective of rank-ordered behavior is just as achievable as when using the more robust linear and logistic regression-style techniques.

Statistical output can be used as key indicators or metrics when comparing the performance of different statistical techniques in the development of a desired solution. These statistical diagnostics can prove extremely important for statisticians and mathematicians, but these metrics can prove quite difficult for business practitioners to review. Business practitioners and executives need metrics and measures that evaluate different options in a very pragmatic manner. As seen previously in the case of gains charts, metrics such as cost per acquisition, cost per order, cost per retained customer, and, ultimately, ROI are among the critical key indicators for executives and business practitioners. All these metrics are simple spreadsheet calculations once the gains/decile charts are produced.

The challenge in communicating meaningful business results to business stakeholders lies in the capability of being able to use historical information from which the desired behavior can actually be validated. In the ideal scenario,

a live historical campaign can be used as both our development and validation sample for the construction of the model. In other words, a previous campaign or business initiative was conducted whereby we are able to determine the responders and non-responders. Ideally, this campaign or effort was conducted using a random sample, which means that the names were promoted randomly rather than based on specific list criteria. The notion of ensuring that a random validation sample is critical to any modeling exercise. If the given sample is random, the results from the validation can be applied to the whole population. Conversely, if the validation sample contains only people under 30, any results and modeling learning from this sample can be only be applied to this particular segment of the population.

Evaluating Solutions

In comparing different techniques, approaches, or models, the key to the comparison is to apply them against the same validation sample and to observe the results. If a given approach, technique, or model is superior to a given solution, then a solution should be observed whereby the trend line or Lorenze curve is steeper. Again, this is applicable for consumer behavior models (Fig. 13.5).

In the above example, it is apparent that the steepest line or slope is model 3, indicating that this solution best rank orders the observations. As rank ordering is our primary measure of success, then model 3 is the optimum solution.

Practitioners are often faced with determining what is the preferred tool for validation in consumer behavior modeling exercises. Is it gains charts/decile charts or an AUC curve? The answer is "it depends," which is just what a statistician would say. Yet, if the decision is to quantify as well as to determine the best model, then the best validation is the AUC curve, as model performance can be directly measured through the KS statistic or Gini coefficient. By contrast, if the decision is to determine the best precise target audience as well as the best model then the gains/decile charts are the best validations since their use allows the reporting of incremental ROI at different decile cutoff levels.

As the world of data science continues to grow, it becomes ever more paramount that there are consistent methodologies when evaluating new data science techniques. Leading-edge practitioners and academia will continue to

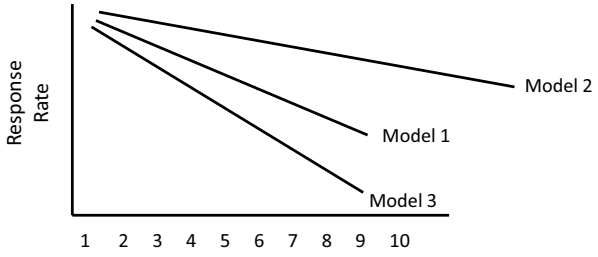


Fig. 13.5 Using Lorenz curves to determine the optimum solution

push the envelope in identifying solutions/software/approaches that will further enhance business solutions. Yet business-minded practitioners in data science must have a robust way of evaluating new techniques. For example, companies are often inundated with new technologies and software that purport to provide superior solutions. These data sets are commonly referred to as research and development (R&D) data sets. Models and solutions have already been developed from these data sets and applied to validation samples in order to provide a performance expectation level.

New technology or software is applied to these data sets in order to determine if and how much incremental value is provided by this new technology.

The same approach in evaluating models is also taken when evaluating new data sources. In this case, the analytical file is recreated using the new appended data. Once again, solutions with and without the new data can be compared in terms of incremental performance. This incremental improvement in performance, as seen by the increased steepness or slope of the Lorenz curve, can then be weighed against the cost of the new data in terms of whether or not this data is economically justifiable.

Applying different approaches, techniques, and new data sources in attempting to arrive at an optimum data science solution should utilize some standard evaluation yardstick. The use of a gains/decile chart or AUC curves are certainly very good examples of the employment of standard evaluation yardsticks. Such measures allow data scientists to truly derive the necessary insights that will assist the successful evolution of the data science industry.

Key Takeaways

- Gains charts and AUC curves are the practitioner's validation tools for consumer behavior models.
- Accuracy rates and loss rates are the only validation option for image recognition and NLP-type models.
- Use AUC curves to mathematically quantify model performance.
- Use decile/gains charts to make decisions based on ROI or some other business metric.
- Both validation options can be used in determining or selecting the best-performing model.

CHAPTER 14

Using RFM as One Targeting Option

Keep in mind that not all data mining solutions require tools with statistical analysis. Although we have focused much of our discussion on building predictive models, the notion of doing something simpler can, in many cases, be equally appropriate. A good example of this is the recency, frequency, and monetary (or FRM) index, which represents one non-statistical method of targeting customers for a given business initiative. Here the analyst can use three key pieces of information:

- Recency since the last purchase.
- Number of purchases in the last year (frequency).
- Average Value of Purchase (Monetary Value)

Let's look at an example (Fig. 14.1).

In this above example, separate indices are calculated for recency, frequency, and average value of purchase. They are calculated by dividing the average value from the entire database against the specific value for a given customer. With an index of 1 being average, values above 1 represent the above-average performers while the reverse is true for values below 1. You will note that the calculation is reversed for recency as low values are good. The index calculations are as follows:

- Recency: $6/3 = 2$
- Frequency $4/3 = 1.33$
- Average Value of Purchase (Monetary): $150/100 = 1.5$

The actual RFM index for Customer A would then be $(2 + 1.33 + 1.5)/3 = 1.61$ where we assume that each behavior (recency, frequency, and monetary behavior) are weighted equally.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_14

Fig. 14.1 Example of indexes for recency, frequency, and monetary value

	Recency- # of months since last purchase	Number of purchases within last year	Average value of purchase
Customer A	3	4	150
Customer Database	6	3	100
Index	2	1.33	1.5

Fig. 14.2 Example of RFM Index using 5 X % matrix

	Recency	Frequency	Monetary	RFM
Customer A	5	5	5	15
Customer B	4	5	3	12
Customer C	5	4	3	12
Customer D	1	2	2	5
Customer E	2	1	2	5
Customer F	1	1	1	3

Another approach is to sort the customers along each behavior (recency, frequency, and behavior) into 5 quintiles. See example given in Fig. 14.2.

In this example, 5 represents the best-performing behavior while 1 represents the worst-performing behavior. In this approach, we just simply sum up the values where the higher the number, the better the overall behavior. In our example above, Customer A is the best customer, with an RFM score of 15, while Customer F is the worst, with an RFM score of 3. Once again, we assume each behavior metric to be weighted equally. Using this approach, you will observe that Customer B and Customer C are good performers, with identical scores of 12, while Customer D and Customer E are lesser-performing, but again with identical scores of 5. There will be many customers with ties in this approach as different combinations of behavior values can end up with the same score.

With RFM, decile reports can be created which rank order or sort customers into groups with decile 1 being the best RFM group and decile 10 being the worst RFM group. See the example given in Fig. 14.3.

In the example shown in Fig. 14.3, customers can then be selected for a variety of different business initiatives.

The use of RFM is straightforward, and it offers a very pragmatic solution as it is able to select records based on prioritizing historical performance behavior.

The limitation of this technique is that it is considered reactive as decisions are based on what has happened in the past. This is not necessarily bad, as the past can often be a reliable indicator of what is to come. In today's rapidly changing world, however, there is the need to be proactive or preemptive as businesses are increasingly seeking decisions for a future in which an increasing

% of Customers	Avg. RFM Index of interval	# of Customers
0-10%	3	10000
10%-20%	2.5	10000
20%-30%	2	10000
30%-40%	1.75	10000
....		
90%-100%	0.3	10000

Fig. 14.3 Example of RFM being used as a targeting tool to select customers

number of variables are uncertain. Analytics that are predictive where past behaviors such as RFM, as well as other behaviors, demographics, and engagements can be related to some future behavior allow organizations to be more prescriptive in their decision-making approach. RFM will always remain as the "quick and dirty" targeting tool, but the use of predictive analytics and predictive models will be the preferred option if solution optimization is the goal.

Many think that the principles of RFM can only be applied in the non-digital world. Nothing could be further from the truth, however, as RFM is essentially about measuring some type of consumer activity or engagement. Yet, in the digital world, these very same principles can be applied in exploring visitor engagement to a website. R would be recency since last visit to the website, F would be the number of times visited to the website, and M might be a proxy for session duration. Using a Digital RFM approach allows marketers to rank order visitors to a website.

Key Takeaways

- RFM is a simple non-statistical approach in targeting existing customers.
- There are two approaches (quintile and index) and both are equally appropriate in targeting existing customers.
- RFM can also be used to rank-order visitors to a website.

CHAPTER 15

The Use of Multivariate Analysis Techniques

The two most commonly used techniques in predictive modeling are linear or logistic regression. The statistics in linear regression predict outcomes with a continuous range of values; by contrast, logistic regression predicts outcomes that are categorical in nature. The most commonly used logistic routines are used to predict "yes/no" behaviors such as response, attrition or credit default. The derived outcome can also be a set of rules such as Chi-Square Activation Interaction Detection (CHAID) rather than a score which is the predicted outcome of either logistic or linear regression.

Both logistic and linear regressions have existed for decades. However, it is only in the last 10 to 15 years that these techniques have become common business practices in marketing as organizations strive to improve their overall targeting efforts. These techniques, as stated earlier, have been widely used for years by companies in their efforts to target customers who are creditworthy.

As with all statistical techniques, the underlying objective is to maximize explained variation and minimize unexplained variation as seen, for example, in this equation to predict response:

Response=	.48
	+.3 × FEMALE
	+.15 × NUMBER IN HOUSEHOLD
	−.00004 × INCOME
	+.07 × NUMBER OF YRS EDUCATION

For now, we can set aside the academic discussion that this equation is linear and attempts to predict a probability function. The academic purists will argue that this equation should be transformed into a logistic function if we want to predict response. As we have discussed in previous chapters, however, the main priority in developing business solutions on consumer behavior is to optimize the solution's ability to rank-order (or differentiate) records based on the

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_15

targeted behavior. In most cases, an accurate estimate of the solution is a distant second priority. However, for now we can assume that this linear function is a good approximation of response.

In the equation laid out above, there are four variables which indicate the following:

- Females are more likely to respond
- People living in larger households are more likely to respond
- People with higher income are less likely to respond
- People with higher levels of education are more likely to respond

In building an equation that maximizes variation, the variation of values among each of these variables best approximates how the response rate values vary. This logic is very similar to what was discussed in relation to correlation analysis. A visual plot of this data, which captures the relationship between age and spending, is provided in Fig. 15.1.

Naturally, the line in the graph represents the predicted outcome from our equation. The points (X) actually represent the observed outcome. In building solutions, the goal is to minimize the distance between the observed points and the line (predicted outcome that is depicted for three points by the letter e). At the same time, the results reveal that there is some trend in the connection between age and spending. The trend (or line) represents a plot of the explained variation. In this example, if spending is our targeted behavior, we observe that age and spending are trending upward; this indicates that age is a good predictor of spending if that is our targeted behavior (since older people spend more). This so-called trend in essence represents the explained variation which, in statistical terms, is often referred to as the R^2. Perfectly explained variation or a perfect equation will yield an R^2 of 1. This implies that the variation of observed response is 100% explained by the equation. In this scenario, every

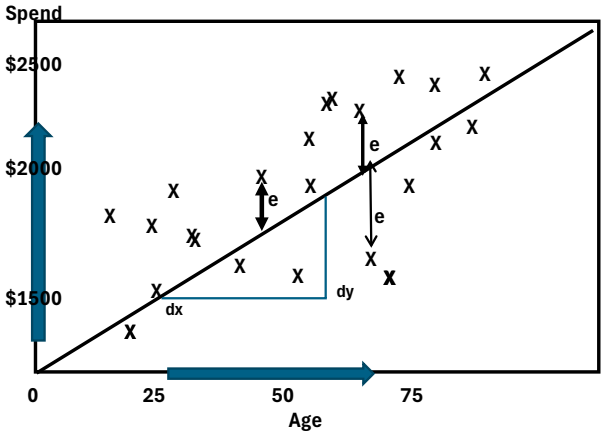


Fig. 15.1 Depiction of best fit line in regression

observed point (X) would be on the straight line. However, in the real world, this is never the case. In fact, the level of unexplained variation is always very high, and this represents a real limitation on the creation of robust solutions. It is highly common to see the R^2 of most equations never exceeding.05. For most data scientists, this represents a real challenge since this high level of residual variation results in practices that call on the use of both science and art.

Some of the newer techniques, such as neural nets, attempt to explain this true random variation, but this variation should be left unexplained. This situation is common among some of the newer more highly advanced techniques, and is often referred to as "overfitting".

With overfitting, these gains charts results are enhanced still further. Once these performance expectations are laid out for the marketing team, objectives and forecasts become quite aggressive and unrealistic and, in most cases, they yield inferior results compared to what was expected. The use of less robust tools such as CHAID or regression analysis in these cases would have delivered solutions that are more easily managed when applied in various business initiatives.

Obviously, this does not mean that these more advanced techniques should never be employed; rather, they, along with other approaches, should always be borne in mind. In fact, from a statistical perspective, these techniques are very powerful, and they have been employed in applications such as fraud detection, particularly in the areas of image and handwriting recognition. The key to using these techniques in business and marketing applications is to have sound validation or holdout samples. The crucial difference within the marketing and business world is that a large portion of the variation in the data is truly random or unexplained. This means that any attempt to explain this "random" variation through the use of more advanced techniques will yield disappointing results. The necessity to use validation samples can never be overemphasized enough in the data science world, especially when large random variation is the norm rather than the exception.

As seen in the section dealing with the creation of the analytical file, the very last stage of the process is the separation of this file into a development and a validation sample. Solutions should always be constructed using the development sample. Then, the solutions' performance should always be measured against the validation or holdout sample; that is, we simply apply the solution to this group of records.

Although it has been argued that is the proper procedure when building solutions, in practice this is not always necessary especially if sample size is a mitigating constraint. The lack of data in building a model would suggest that the development and validation samples should be combined in order to have enough information to develop a proper model. Although there is some risk in developing models without a validation sample, this risk is mitigated with the less robust techniques such as CHAID and regression analysis. To an experienced eye, minimal difference is observed in results when the solution is applied to a validation or development sample, particularly if the proper discipline has

been exercised during model development. If the model and its variables are statistically sound, then it is expected that so too will be the results. These simpler techniques are only producing results from that portion of the total variation that is explainable; in most cases, that portion represent a very small part of the total error. Hence, simpler techniques such as CHAID and regression are sufficient for developing solutions to explain this small portion of the total variation. However, this explainable error must be duplicated when validating results that are similar to what occurred in development. With the more robust techniques, some of the model's predictive capability is based on being able to explain variation which is truly random or unexplained. This attempt to "explain the unexplained" is not replicable in a validation sample and drastically different results between development and validation are expected if a more robust and complex model were not properly trained. In my experience, when models have been properly trained with these more robust techniques, the results are indeed similar to those produced by CHAID and regression.

In the world of marketing, situations will rarely arise that allow a large portion of the variation (i.e. customer behavior) to be explained. In other words, techniques like neural nets and machine learning seldom outperform the more traditional styles of regression techniques. However, the small portion that can be explained results in significant increases in performance. Because these solutions are applied to hundreds of thousands of customers, these simple models can achieve a tremendous lift in each customer behavior and, ultimately, add hundreds of thousands of dollars in revenue to a given campaign.

Understanding that this small explainable error is the norm in data science can help explain why linear or logistic regression can be used to predict binary outcomes. While from the statistician's purist perspective, this is heresy, from a data scientist's perspective, it is simply a matter of what works and what doesn't.

Binary outcomes, such as defection, credit loss, or response rate, are essentially "yes/no" outcomes that are being predicted. Did a person respond? Did a person go into default? Did a person cancel?

In the world of data science, the rudimentary use of statistics will often suffice and be preferable to the more complex approach. As discussed throughout this book, the real key to building stronger solutions is a data environment that offers much data that is reliable and provides the flexibility to manipulate this data information into more meaningful fields.

Because of the binary nature of outcomes such as those mentioned above, these models should be producing a probability function. In other words, the outcome should produce results that are between 0 and 1. Academic purists will argue that using linear regression to estimate binary outcomes is an incorrect approach since the predicted outcomes from a linear regression routine will have values between 0 and 1. From the practitioner's perspective, however, what matters is that the solution provides significant performance lift when compared to other alternatives. As previously mentioned, performance lift is the practitioner's primary focus. Of course, it would be nice to have a solution that is also accurate in predicting actual response rate (i.e. the difference

between the predicted response rate and actual response is minimal), but this is not mandatory. Stakeholders and data scientists desire solutions that can best differentiate or rank-order groups in their differences from the average. The example in Fig. 15.2 illustrates this focus.

In this example, Scenario 2 produces a model that more accurately predicts response. However, which scenario would the business stakeholder select as a solution? Despite its obvious inability to accurately predict response, Scenario 1 still yields a much higher lift and would be the preferred choice of the business. The ability to produce a solution that offers a significant lift in performance is the preferred objective over one that provides more accurate estimates of response rate.

If lift is the primary consideration, then there is no need to have a mathematical routine, such as logistic regression, which delivers probability estimates. Once again, given the small amount of variation that can be explained, in most cases linear regression will yield solutions that are often very similar to logistic regression when considered from the perspective of lift.

Why, then, should logistic regression be considered at all in a business setting? The practical response is that models are often built as part of an overall solution, such as campaign optimization or profitability. In these cases, the predicted outcomes represent probability estimates that may be used as one metric in building the overall solution. Here's an example:

$$\text{Profitability} = \text{P}\left(\text{Response}\right) \times \text{P}\left(\text{Approval}\right) \times \text{Spend}\left(1 - \text{P}\left(\text{Cancel}\right)\right)$$

In this example, profitability is determined by a person's likelihood to both respond to, and be approved for, this offer and the applicant's level of spending for a given period, as well as the likelihood of being a customer at the time of the offer.

Converting models into probability functions is not very difficult. Once users have identified the key variables in the final linear regression, they simply input the same variables into a logistic function readily available in many software packages. In fact, in building logistic or probability models, this may be

	Scenario 1	Scenario 1	Scenario 2	Scenario2
	Predicted Response Rate	Actual Response Rate	Predicted Response Rate	Actual Response Rate
Top 10% of Model	6%	10%	4%	4.50%
Random Control Cell	3%	5%	3%	3.50%

Fig. 15.2 Comparing predicted estimates vs. observed estimates

the preferred approach especially if the central processing unit (the CPU) and time is an issue. It is far quicker to determine the relevant variables and statistical variables from linear regression. For the most part, the selection and identification of these variables is almost the same as when using linear regression or logistic regression as the sole routine in arriving at a solution. Given this finding, and the higher CPU requirements of logistic regression, it is preferable to use linear regression for variable discovery of the model exercise. Then, once the model variables are finalized, they are transformed into a logistic probability function as the last step in model development. In situations where solutions are built that require the integration of several models, the combination of linear regression with logistic regression provides analysts with the appropriate tools to yield the best solution with quick turnaround time. This then allows analysts the ability to turn their focus to the specifics of the solution an organization requires.

The Impact of Multicollinearity

Statistics play an extremely valuable role in optimizing the decision-making process. Understanding what statistics can do in relation to a business objective is critical to this process. This is certainly the case when marketers are asked to create a target group or list of customers and at the same time identify the key characteristics defining that group.

For instance, when a list of customers is generated, the objective is to obtain the best group or list of customers. Regression analysis routines perform very well in achieving this objective. The use of these routines helps to identify the characteristics that will best predict a given outcome. One real advantage of these techniques is that they also consider multicollinearity.

What is multicollinearity? A good practical example is the notion that education usually leads to higher income. In another example, such as response or attrition, the high multicollinearity of both age and income would end up causing both variables to exhibit the same trend on response or attrition. In another example, we might find that gender and tenure (length of time as a customer) have no relationship with each other. In conducting any analysis on response or attrition, we cannot assume a priori that the two variables (gender and tenure) would have the same relationship on response or attrition.

In mathematical terms, multicollinearity represents the level or amount of interaction between the independent variables in a given model equation. This means that the weights or coefficients of the final variables reflect this interaction between multiple variables and, at the same time, are statistically significant in the overall equation. Through regression analysis techniques, many variables will be eliminated because of multicollinearity with the final model variables, but they may still be statistically significant in a univariate fashion as seen in correlation analysis routines. This type of process is critical for optimizing the predicted outcome. The key outcome in this case is a score that is the key measure used to generate lists or group of customers.

	Years of Education	Income
Correlation Coefficient	0.11	0.12
Confidence Interval	99.0%	99.5%

Fig. 15.3 Example of correlation results of education and income vs. response

When attempting to identify the key characteristics that define a given group, using the characteristics or variables from a regression routine may be less than optimal. For example, Fig. 15.3 presents a case in which the goal is to optimize response rates.

Both variables (education and income) are statistically significant with response rate in a univariate fashion (i.e. each variable without the impact of multicollinearity vs. response rate). Both education and income exhibit a positive trend with response rate.

Yet, the regression equation below, however, only includes income:

$$\text{Response} = 0.50 + 0.00001^{*}\ \text{income}$$

Here the equation has retained only one of the two statistically significant variables (income). The high multicollinearity between income and education has caused education to be eliminated. However, in identifying key customer triggers or behaviors for response, marketers would want to know that both education and income are key customer characteristics in the target group. A more dangerous scenario would be the case in which both variables (income and education) are included in a model, but where one of them has a negative sign. In this case, the negative variable, which should have a positive sign based on its positive correlation with response, needs to be either replaced with another variable or eliminated from the equation.

Similarity Between Logistic Regression and Multiple Regression

The similarity of logistic regression to multiple regression can be easily observed in the actual logistic equation, which is $1/1 - \exp(\text{Y PRED})$, where YPRED is the actual output or score of the multiple regression equation above. Logistic regression simply transforms the multiple regression routine in order to ensure that its outputs are probabilities ranging from 0 to 1.

Other techniques, such as discriminant analysis and support vector machines (SVM), are classification techniques that create algorithms which best differentiate a group of classes. The more commonly used tool of the two is SVM since it has no assumptions regarding normality.

The Use of Recommender Engines

For many organizations today, the issue is not one of an information deficit but rather one of information abundance. In this environment of abundance, the challenge becomes one of priority by being able to offer the right product or service for the consumer at that point in time. This can be very daunting given the many products and/or services of many organizations alongside all this information. So, what can organizations do?

Taking their cue from such organizations as Netflix or Amazon, recommender engines have been increasingly used by organizations, particularly those with a large suite of products and/or services. Without delving too deeply into the math, the underlying concept is to explore the existing pattern of purchased products and services where the engine then offers the next suite of products/services to the customer. There are many options that are available to the data scientist. They typically conduct analytics at two levels, with the first level being at the product/service level, often referred to as the item level. In a way, this option is conducting massive correlations between products in order to detect similarity purchase patterns between products. Several different options are readily available in Python, and these include such options as Jaccard similarity or Cosine similarity.

The second level is at the customer level or user level, where clustering techniques are used to identify distinct customer groups who would have similar purchase patterns regarding certain products and services. This referred to as the user-based approach.

Some organizations may use a hybrid approach, which is the combination of both the item-based and the user-based solutions. In these cases, each item would have both a user-based and an item-based score. By combining both scores together, the analyst can create a composite score and then rank each item or product based on that composite score.

Obviously, these types of tools make sense when there is a vast array of product/services, such as occurs within the retail industry. But what about those scenarios when products and services are limited in scope? The use of customized predictive models can then be extremely useful in predicting broad courses of action. For instance, within a financial institution, these broad product/service areas might be as follows:

- Credit card
- Mortgage
- Line of credit
- Loans
- Wealth/investment

Predictive models could be developed in two broad areas, where the first activity represents the likelihood of purchasing that product or service while the second activity represents the likelihood of retaining that product or

service. With model scores built for each activity, the bank could determine the next best likely outcome.

One advantage of predictive models in a low-cardinality environment (few products and/or services) is the simplicity of the process and our ability to better understand a given solution. The very definition of "predictive" is better reflected in the fact that we have both a pre-period (independent variables) and a post-period (the dependent or target variable).

One advantage of recommender engines over predictive models is that they function better in a high-cardinality environment (one involving many products and/or services). They still yield predictive estimates, yet their estimates are based on historical information and not the pre-period vs. post-period design of information, as described above in developing predictive models. Because of the issues related to sparsity and cardinality, the use of recommender engines is better suited analytically to exclude sparse-type products, and to focus on those products which have enough purchase data. This makes it a more suitable approach than predictive models using a pre- vs. post- window.

KNN (Nearest Neighbor Modeling)

One of the more fascinating developments in data science and its evolution within Big Data is the use of KNN (Nearest Neighbor Modeling) as a technique for not only modeling, but to also provide estimates for variables where there are missing values. In modeling, it works as follows. Through techniques such as correlation and regression analysis, one identifies the key variables in the given exercise. Suppose we are developing a response model and that, through this analysis, we identify the following key variables that most impact response are income, education, and household size. The user then decides the number of nearest neighbors for each record. Let's say that number is 3. Figure 15.4 gives a very simple schematic of what this might look like.

In the above example, the record of interest, denoted by the star, determines its three closest neighbor (brown circles) based on the fact that these three records give the minimum difference in income, education, and household size from the star's record values of income, education, and household size. We also

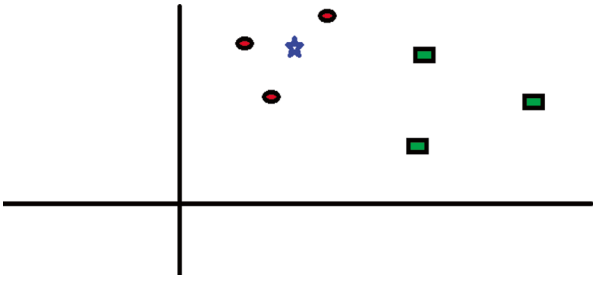


Fig. 15.4 Simplified example of a KNN model in action

have the actual target response variable behavior for each of the brown records. If two of the three brown records responded, then the predicted response rate for the star record is.66. We then determine the actual response rate behavior of the star record, which would be 0 for non-response and 1 for response. Each record now has a predicted response score and observed response.

We are now in a position to validate our observed result versus what we have predicted. In Fig. 15.5, we can see that the KNN model validates quite nicely when comparing predicted response scores to observed response scores. The data scientist also has the latitude of playing with different numbers of nearest neighbors alongside trying to utilize different input variables. At the end of the day, however, it is about arriving at a solution that produces the best model validation results.

But besides modeling, KNN can also be used to impute values for variables where there are missing values. Let's explore this more closely. Once again, let's say we want to impute market research information related to customer attitudes towards government spending. We have 100,000 customers, and only 2000 customers with the requisite market research information related to government spending with missing values for the remaining 98,000 customers. For each of the 98,000 records, we find the nearest three records from the 2000 market research sample, assuming a KNN factor of 3. Once again, we would need to identify the key database variables where we could calculate the minimum distance between the missing value record and the three nearest market research records. From these market research records, we then calculate the average of the respondent information pertaining to government spending, which is then assigned to the missing value record.

The challenge again in this case is to determine the appropriate KNN factor and the key database variables in which to calculate the distance information. Once again, we could find out what database variable in the market research sample is highly correlated with attitudes on government spend. Suppose, in this example, that variable is income. We then use different KNN factors to determine the optimal factor based on validation.

	KNN =3				
	nearest neighbour 1 response behaviour	**nearest neighbour 2 response behaviour**	**nearest neighbour 3 response behaviour**	**total predicted response for record1**	**total observed response**
record1	0	0	1	0.33	0
record2	0	0	0	0.00	0
record3	1	1	0	0.67	1
record4	1	0	1	0.67	1
record5	1	1	1	1.00	1
record6	0	1	0	0.33	0
record7	0	1	1	0.67	1
record8	1	0	0	0.33	0
record9	0	0	1	0.33	0

Fig. 15.5 Sample records using KNN in estimating response behavior

	KNN =3			
	nearest neighbour 1 prefers government spend	nearest neighbour 2 prefers government spend	nearest neighbour 3 prefers government spend	prefers government spend for non market research record
record1	0	0	1	0
record2	0	0	0	0
record3	1	1	0	1
record4	1	0	1	1
record5	1	1	1	1
record6	0	1	0	0
record7	0	1	1	1
record8	1	0	0	0
record9	0	0	1	0

Fig. 15.6 Sample records using KNN in estimating preference for government spend

Let's look at an example. Suppose we use a KNN factor of 3, which calculates the minimal distance in income between the three market research records and the record not in the market research sample. Note that income will be on both the market research record and the non-market research record. We calculate the average information related to government spending of the three nearest market research records and then impute the majority opinion to the non-market research record. Let's look at an example (Fig. 15.6).

In this example, each non-market research record (9 records) are assigned either a 0 (do not prefer) or a 1 (do prefer). This assignment is based on the ratio of the preferences from the nearest neighbors. If it is more than 50%, the record is assigned 1; otherwise it is assigned 0.

Key Takeaways

- Overfitting is more of a concern for the more advanced techniques like neural nets.
- Simple techniques work quite well, particularly in an environment where unexplained variance or noise is quite large.
- Most statistics, including predictive models, are about minimizing error, with R^2 being the most common metric used to assess multiple regression models.
- For the most part, logistic regression and multiple regression perform equally if rank-ordering of the target variable is the modeling objective.
- Multicollinearity is a critical factor in regression modeling routines and needs to be addressed when producing models.
- Recommender engines are mathematical techniques used to determine what the consumer will prefer next. User-based and item-based approaches are used to create hybrid scores that are used to select the next product or service.
- KNN (Nearest Neighbor Modeling) can be used as a predictive modeling technique, or used to impute values on records with missing variable values.

CHAPTER 16

Tracking & Measuring

Data science is much more than just technology. It is a step-by-step process resulting in the development and application of a solution. Yet, at the same time, this process should always yield learning which can potentially be utilized for future campaigns. The development and application of solutions, as well as the acquisition of new learning, provide the means for continuous business improvement, which is the long-term goal. It is in this stage that learning and insights are provided for future business initiatives. In fact, the leading-edge organizations devote much of their resources and time within this area. Why? One of the keys to the long-term success of data science is the creation of an environment which continually optimizes learning to the broadest audience possible. Having the ability to measure and evaluate results as part of an ongoing process provides the insights that are required for continued learning. However, it is equally important that these results are effectively communicated such that the information and insights are distributed to a wide audience. The empowerment of more people through increased access to these results simply means that more minds can be focused on building solutions, rather than just a select few.

Consider how measurement and evaluation fit into the overall data science process. The first stage (Identification of the Business Problem/Issue) deals with the collection and gathering of information which will enable a data scientist to identify key issues and challenges. The ability to measure and evaluate results from recent business initiatives and campaigns enables the data scientist to gather the most up-to-date information in helping to formulate these key issues and challenges that continue to arise throughout this ongoing learning process.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_16

Measuring ROI: Getting Started with the Data

Understanding what is to be measured within a business initiative can be the most significant hurdle to effective ROI measurement. Situations where measurement reports are produced, but do not yield the required results often occur in marketing. Resolving or mitigating these types of situations requires an approach that really looks at addressing all the needs of those stakeholders who are involved in this type of process.

Presuming that these measurement objectives and stakeholder needs have been identified, the next issue to tackle is: how will this be done? It is this question which requires us to look at the data and to then determine if the identified measurement objectives can be met within the existing data environment.

For example, how is the appropriate testing matrix set up for a given business initiative or marketing campaign? What kind of learning does the marketer want to obtain, and how do they evaluate the performance of a given initiative or campaign? For instance, if a targeting tool is being deployed within a campaign, results need to be obtained regarding its performance. However, it is the amount and complexity of the required learning which will dictate the design of the measurement and tracking system. For example, the simplest kind of tracking is to merely determine the performance of a particular initiative. This initiative could be anything, such as an event, campaign, type of offer, type of communication, targeting tool, etc. In this case, the tracking requirements are quite simple in the sense that two cell codes/groups are created. The first cell code or group is the control cell in which there is no occurrence of the initiative, whereas the second group or test cell is impacted by the initiative. This is rather easy if the marketer's interest lies within only one piece of learning from a campaign. Yet, in practice, marketers are always trying to obtain as much learning as possible. In setting up a measurement matrix, complexity is not the issue; the real issue is costs. For each given initiative, test cells need to be set up. Suppose that a marketer wants to evaluate 4 different offers, along with a control offer. In this case, a total of 5 cells would need to be created.

Now suppose that the marketer wanted to test 4 different offers, along with 4 different communication pieces. More importantly, suppose they wanted to test the interaction of offer and communication. Consequently, they would need to create 16 separate cells (4 offers × 4 communication types), in which offer type 1 is the control offer and communication piece A is the control communication piece. Figure 16.1 is a schematic of this tracking matrix with live promotion quantities.

In any tracking scenario, there is always one cell which the marketing team has to select as its winning cell. Marketers will place many of the names here since this is the cell which is expected to generate the largest return on investment (ROI). In the example here, the winning cell is D4, while the other 15 cells essentially represent the investment costs of promoting 75,000 names or 5000 names per cell. This can be considered a significant investment since 20%

Offer Type	Communication Piece			
	A (Control)	B	C	D
1 (Control)	5,000	5,000	5,000	5,000
2	5,000	5,000	5,000	5,000
3	5,000	5,000	5,000	5,000
4	5,000	5,000	5,000	**200,000**

Fig. 16.1 Example of marketing matrix for measurement

Offer Type	Communication Piece			
	A (Control)	B	C	D
1 (Control)	X	X	X	X
2	X	X	X	X
3	X	X	X	X
4	X	X	X	X
Region				
Maritimes	X			
PQ	X			
Ontario	X			
Rest of Canada	X			
Targeting				
Segment 1	X			
Segment 2	X			
Segment 3	X			
Segment 4	X			

Obs. Each region cell and targeting cell get control communication and control offer

Fig. 16.2 Example of marketing matrix for measurement (more complex)

(75,000/275,000) of the promotion quantity is devoted to learning. At $1.00 per piece, this can translate to $75,000 in investment devoted to learning.

Suppose the marketer wanted to introduce other factors into the mix, such as region (4 groups) and targeting (4 groups). In such circumstances, the tracking matrix in the previous example would then explode into a 256-cell matrix (4 offers × 4 communications × 4 regions × 4 target groups). Obviously, this soon becomes unwieldy.

This is where the use of statistics such as conjoint analysis or Analysis of Variance (ANOVA) can be used to reduce the number of test cells to a more manageable number. These techniques are useful in providing some science regarding the interaction between factors. Supposing that conjoint analysis/ANOVA demonstrates that there is no interaction of targeting and region with the other factors. Then, the following matrix could be produced (Fig. 16.2).

Using statistical analysis, the number of cells has been reduced from 256 to 24 cells. This is another illustration of the important role played by statistics within database marketing programs. Yet, in this scenario, the objective is not the creation of targeting tools but rather to determine what cell/groups need

to be tested within a given marketing campaign. Bear in mind that we are also assuming that there is no impact between offer type and communication type on region and/or targeting based on statistical analysis such as ANOVA and/or conjoint analysis. In other words, we are exploring the impact of region and targeting independently of offer type and communication type.

Digital marketing provides the capability to explore all kinds of different test scenarios as the cost to create different page views and email types is negligible. Although one now has the capacity and resources to create these complex measurement scenarios, the limiting factor becomes the human capacity to derive meaningful learning from the multitude of tests that could be designed for a particular marketing initiative. Using this science to reduce factors for testing accelerates the level of marketing analysis to a new level where analyses reveal more meaningful insights and in a timelier manner.

Another issue which often arises when designing a testing matrix is deciding which sample size to allot to each cell. There are a variety of factors that impact sample size. Examples of these factors include the overall performance rate, the error range, and the confidence level, all of which influence sample size. The actual formula can be written as:

$$\frac{(\text{Confidence level})*(\text{Confidence Level})*(\text{Performance Rate})*(1-\text{Performance Rate})}{\text{Error Range}*\text{Error Rate}}$$

Here are some examples to obtain clearer insight into how this formula works. Assume a response rate of 1% (performance rate), a confidence level of 95% which translates to a statistical Z score of 1.96, and the fact that the results should be statistically significant. If given initiatives are 0.2% (error range) different from the mean, then the minimum required sample size is simply:

$$\frac{1.96*1.96*(.01)*(.99)}{.002*.002}=9508$$

If the allowable range of error is increased to 0.4%, then intuitively one would expect the minimum required sample size to decrease. In fact, it does decrease, to 2377. Increases to the confidence level will always increase the sample size. For instance, increasing the confidence level from 95% to 99%, or the ZSCORE from 1.96 to 2.58, we see that the sample size increases from 9508 to 16,475. Meanwhile, a performance rate of 50% always yields the largest sample size since any other combination of performance rate and its complement (1-performance rate) will always yield a smaller number in the numerator.

This formula can be plugged into any spreadsheet application. It has tremendous utility to the marketer in terms of its ability to provide a range of required sample sizes based on the differing assumptions of performance rate, range of error and confidence level.

Suppose a marketer wants to be more granular in their results performance by determining the optimization of a given business objective. This scenario is especially applicable to targeting tools that are used within a marketing campaign. With a control or random cell set aside and then used in comparison to the targeted group, a marketer can quickly ascertain if the model is effective by observing if the results are superior to the control group. But is it optimal? Is the targeting performing as was observed when the tool was built? This is where a comparison between when the model was built and its current application needs to be created. By checking the performance results of the model when it was initially developed, the marketer then has a benchmark against which to compare subsequent targeting application results.

For instance, suppose a response model is built whereby customers are ranked into 10 deciles, with decile 1 being the highest scored and decile10 being the worst scored. The observed response rate from the validation sample for model development is then reported for each decile and plotted as a line or curve. This line or curve is often referred to as a Lorenz curve. The objective of any model or targeting is to maximize the slope of the Lorenz curve. A flat line is complete failure while a vertical line represents perfection.

Once this model is applied to a current campaign, the exercise can be conducted and a Lorenz curve can be created for the current campaign. This provides a means of determining optimality through a comparison of the slopes of the two Lorenz curves. If the slopes are identical, then the model is performing optimally. By contrast, if the slope of the campaign is less than the slope for model development, then the model is not performing optimally. The key business decision in this type of situation, though, is to determine whether a company can live with this sub-optimal performance or invest in the redevelopment of a new model.

Here are a couple of examples to better illustrate this concept. Figure 16.3 gives an example of two Lorenz curves, the first one reporting the results during model development (Validation) and the other reporting the results after the model was applied for a campaign (Campaign).

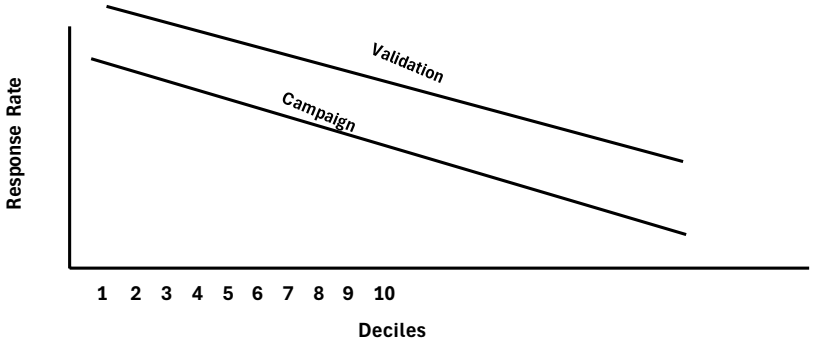


Fig. 16.3 Example of model evaluation in a given campaign

Even though the campaign's overall response rate is performing much worse than what was observed during model development, the targeting tool is performing very effectively as its ability to rank order response rate has remain unchanged between model development (validation) and its current application (Campaign). Since the slope of the two lines is identical, it can be concluded that the model's current performance is optimal. In this case, it was critical to evaluate the model's performance in a granular manner since the initial instinct of the business team might be to blame the model for the poor performance of the overall campaign.

Figure 16.4 offers another example of reviewing a model's performance during its current application.

Observe that the overall campaign is performing much better than what was observed during model development. In this case, the first inclination might be to credit some of the campaign's success to the model when, in fact, the model had provided no contribution toward its success. This is because there is no rank ordering of response rate by decile, as demonstrated by the flat line Campaign Lorenz curve. The slope of the Campaign Lorenz curve in this case is almost non-existent. The business decision in this case would be relatively simple: a new model needs to be developed.

These are just some examples of what one might consider in designing a measurement and tracking system. The key in the creation of any measurement and tracking system from a tactical standpoint is to establish priorities and objectives and plan accordingly. Yet, this planning can be irrelevant if the practitioner does not properly understand how to use the data within the existing information environment and, more importantly, how to use it in a cost-effective manner.

Another way to measure performance of a model is through what is commonly referred to as the AUC curve (Fig. 16.5).

In the example above, which is a risk model to target homeowner losses, the analyst is tracking the impact of a predictive analytics model (model produced by our company, BFG) vs the status quo (current pricing). The current way of pricing (blue line) does a very adequate job in identifying losses when

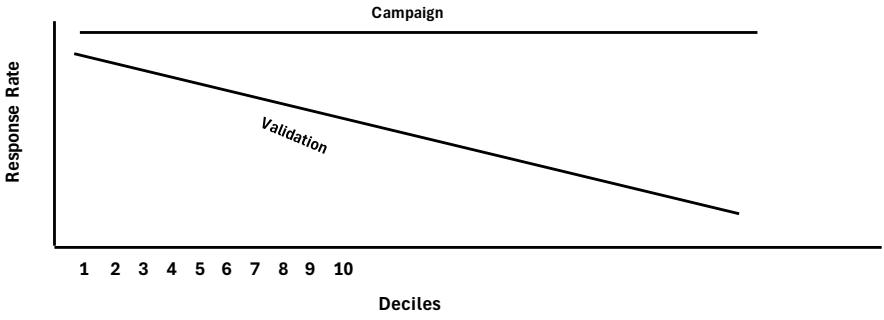


Fig. 16.4 Example of model not performing

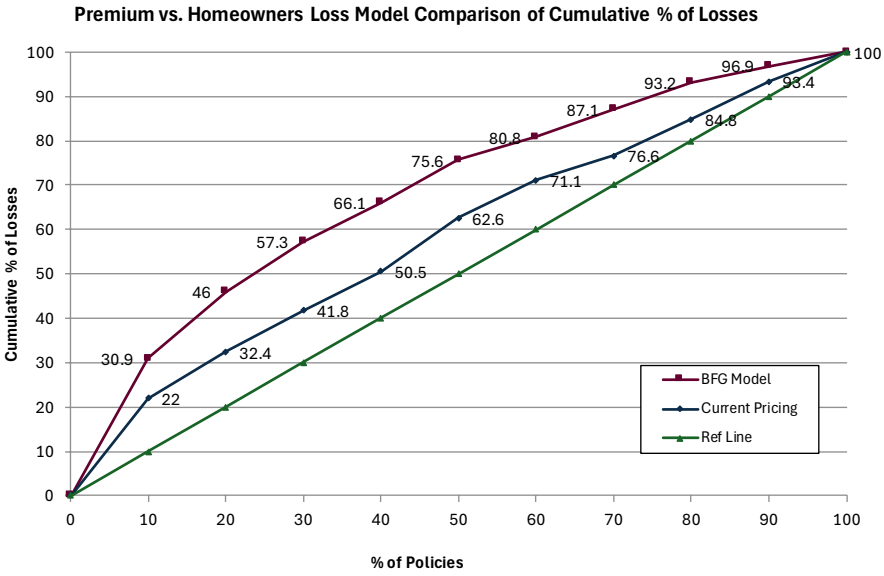


Fig. 16.5 Premium vs. homeowner's loss model comparison of cumulative % of losses

compared to doing nothing, which is the straight green line. Yet, the predictive model (BFG-maroon line) does significantly better, as depicted by the area highlighted between the BFG model curve (maroon line) and the current pricing curve (blue line).

Creating Match Keys and the Challenges of Marketing Attribution in a Digital World

Measurement has never been an easy task and has become more complex in our increasingly digitized world. As with most exercises discussed in this book, creating the analytical file serves as the foundation in any measurement exercise which, of course, must include the concept of match keys in linking data together. Let's explore this a little further.

In identifying the objective function or target variable for a given analysis, one must often match files together to create the target variable. If one is matching existing customer files, account/customer ID, email, and phone numbers all provide easy options in trying to obtain a target variable. However, this process is not so easy when you are trying to conduct analysis for acquisition programs. In such cases, the match keys that are typically available need to be developed from name and address information. In matching a responder file back to a promotion file, the typical methodology in creating a match key comprises postal code, last name and 1st digit of first name. This is what we call a "looser" type of match key as if we had R. Boire 4628 Mayfair Ave H4B2E5 on the promotion file, then the looser match key would be H4B2E5BOIRER. If the responder file

had Rick Boire 4628 Mayfair Ave. H4B2E5 on the promoted file, then the loose match key of H4B2E5BOIRER would consider that there is a match between the responder record and the promotion record. If we had used full name and address, which is considered a tight match key, then Richard Boire would be considered a non-responder. But the utilization of the appropriate match keys in matching files is just one challenge in measurement; the more significant one is marketing attribution, especially in an increasingly digital environment.

Digital marketing can be a challenging phenomenon. Not only does there exist the challenge of navigating across a myriad of platforms to create the appropriate marketing communication to the consumer, but the bigger challenge is also: how do we effectively measure the impact of the campaign?

These challenges also existed prior to the internet. Marketing channels comprising direct mail, outbound telemarketing, sales force, TV, radio, magazines, and newspapers were all used to promote the services and products of the company. If a consumer received a direct mail package alongside some TV and radio advertising, the burgeoning question was how to allocate an actual consumer purchase across the three distinct channels of the marketing campaign. How do we attribute the consumer spend back to the appropriate marketing channel? Defined as marketing attribution, our goal is to effectively attribute the right spend to the right marketing channel, thereby allowing us to allocate appropriate weights for each channel. Let's explore this more deeply.

During the mid-1980s, in my early career as an analytics professional at Reader's Digest, the issue of marketing attribution was nonexistent. All marketing execution was conducted through direct mail. The direct mail campaign had specific campaign codes and responder pieces that directly attributed the spending of that consumer to that channel if he or she responded to that campaign. The attribution was simple here: 100% attribution to the direct mail channel. This approach was consistent from campaign to campaign, thereby providing a certain degree of consistency for the results. Yet it was flawed as the Digest always had radio, TV, and sometimes print ads, which were never considered as part of the marketing mix. To assume that these channels had no impact on sales was clearly to overstate the impact of direct mail on sales. However, in the 1980s, that was the approach.

With the introduction of the internet and social media, a plethora of new channels, such as programmatic advertising, email, text messaging, are now deployed against consumers. The growing emergence of digital marketing in the business landscape has simply underscored the need for a more disciplined approach to marketing attribution. In the early 2000s, the approaches were simple, with one being to attribute the success of a sale to the last marketing channel or to attribute the success to the first marketing channel. This soon evolved into attribution of the sales equally to all marketing channels communicating to the customer during the campaign period. For example, if there were three channels (direct mail, programmatic advertising, and email), and a customer purchased a $100 pair of tennis shoes, then each channel would be attributed 1/3 of the success, or $33.

These approaches above do attempt to produce some rigor in the process rather than the 100% assignment to direct mail for all campaigns. Yet they are still somewhat intuitively based. In other words, could we arrive at something more analytically based? The answer is "yes," which is the marketing mix model where an actual regression model is developed to assign the appropriate weights to each market. This works as follows:

1. Collect the data (i.e. records of interest) which in this case are the campaign period/event
2. Determine what is incremental sales due to marketing for each campaign event. This becomes your target variable
3. Determine the marketing channels used to promote customers during that campaign event/period
4. Develop a regression model predicting incremental sales due to marketing where the independent variables are the specific marketing channels. See example below:
   - Incremental sales = A + 500 × Direct mail +300 × programmatic advertising + 200 × Email
   - Normalize the channel coefficients to %s:
     - Direct mail = 50% (500/1000)
     - Programmatic advertising = 30% (300/1000)
     - Email = 20% (200/1000)

This would imply that the marketing budget should contain 50% direct mail, 30% programmatic advertising, and 20% email. Obviously, there is a large volume of work in creating the analytical file that would be required in building the model.

Another concept in marketing attribution is to determine if the marketing campaign even had the desired impact. Before doing this, however, it is important to understand the overall consumer purchase behavior. In other words, what is the regular purchase period of the consumer within that company and industry? For example, we know that the purchase cycle in buying groceries is going to be much shorter than the purchase of car tires. But how do we determine this? The first approach is just based on the business and domain knowledge of the key business stakeholders. But another approach is more analytically oriented. Let's explore this in more detail.

The underlying question to ask when determining the purchase cycle is to explore the purchase activity of the consumer over a series of months. Figure 16.6 looks at the cumulative % of customers that are active within a given period.

In Fig. 16.6, we plot the cumulative % of customers that are active in 1 month, active in 2 months, etc., up to and including active in a 6-month period. Each line expresses the cumulative % of customers that are active within a cumulative period. From the simple chart above, we observe that this trend increases sharply upwards, but then flattens. It is this point (the elbow point) which is expressed as the average purchase period. For company A, that point is approximately 3 months, while for company B, it is 5 months.

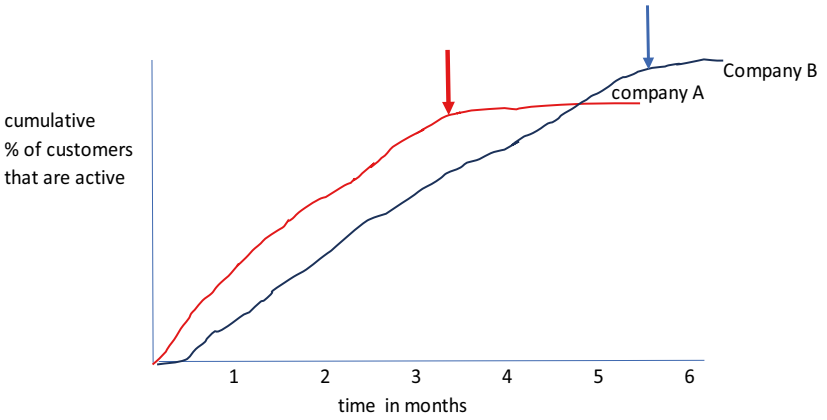


Fig. 16.6 Activity rates by time in months

Fig. 16.7 Analyzing spend behaviours during a marketing campaign

Grocery Campaign Jan.20-27 Average spend per customer		
Pre (Jan.12-19	During(Jan.20-27)	Post(Jan.28-Feb.4)
200	190	210

Once we define our average consumer purchase period, that period becomes the interval of time over which to observe a consumer purchase because of marketing activity. Hence the length of time in measuring any campaign performance should be the same as the average consumer spend period. It is here that the concept of pre-, during, and post- periods come into play. Figure 16.7 is a simplistic illustration of how we might measure incremental spend due to marketing activity.

Note that, in the grocery example, the average consumer purchase period is one week, which is the length of the campaign period (Jan. 20–27). But to look at whether or not the campaign achieved incremental customer spend, we need to have a pre- period, which is the same length of time but occurs just prior to the campaign, as well as a post- period, which covers the same length of time after the end of the campaign. In the given example, marketing had no impact on customer spend. If we had the following results below listed in Fig. 16.8, this would indicate that the campaign was successful and was sustainable, as depicted by the customer spending number in the post- period.

In Fig. 16.9, we see that the grocery spend was increased because of the marketing campaign, but that it was not sustained, as seen by the drop-off in spend in the post- period.

In all three cases, we are looking at overall marketing spend. We could look at spend by channel, however, by simply using the marketing mix model in apportioning out the impact across each channel. Suppose we had the following marketing mix model as observed above: Incremental sales = A + 500 × Direct mail + 300 × programmatic advertising + 200 × Email.

Grocery Campaign Jan.20-27 Average spend per customer		
Pre (Jan.12-19	During(Jan.20-27)	Post(Jan.28-Feb.4)
200	250	260

Fig. 16.8 Analyzing spend behaviours during a marketing campaign

Fig. 16.9 Analyzing spend behaviours during a marketing campaign

Grocery Campaign Jan.20-27 Average spend per customer		
Pre (Jan.12-19	During(Jan.20-27)	Post(Jan.28-Feb.4)
200	250	210

In such a case, direct mail would account for 50%, programmatic advertising for 30%, and 20% for email. If we were to produce a table for email alone, then we would simply adjust these numbers to 20% of the total.

As one can surmise, marketing attribution and assessing incremental spend are key issues in today's increasing digital environment. The approach here is not an exact science, but a pragmatic way of looking at marketing attribution. As we become more acclimatized to a multi-channel environment, accountability will become more predominant and effective measurement will be the real key to success in any marketing endeavor.

Key Takeaways

- What gets measured gets done.
- Data Scientist needs to determine the right data and information in developing a measurement template.
- Both Gains Charts/Decile Charts and AUC curves can determine level of performance with AUC curves better able to quantify performance while gains/charts are better at determining the precise point to make your business decision.
- The creation of match keys is important in the integration of files to create an analytical file.
- Marketing attribution is a key business challenge in an ever-increasing digital world with the use of marketing mix models as the best practice approach in assigning success to specific marketing channels.

CHAPTER 17

Implementation

With the solution complete, the next step is to action it within some business initiative. In some cases, these initiatives could be non-marketing-related. For example, the development of credit-risk models could be applied within the operations area, whereby the actual output of these models to the account representative are risk segments along with their specified courses of action.

In applying solutions, the most important consideration is to ensure that the solution is being applied correctly. Initially, this implies that data quality checks on several records would be conducted to ensure the solution has integrity on these records.

Another consideration is to ensure that the information environment has not changed substantially between the time of the developed solution and the time of its application.

This can be done by creating frequency distributions of key elements within the solution. A model could be examined by comparing the score distribution ranges between time of development and time of implementation (Fig. 17.1).

In this example, the score ranges have changed quite drastically between the time of development and the period of the most current application. Forging ahead with this solution without understanding these score discrepancies is an invitation for failure. Investigation and analysis needs to be conducted on the database in order to clearly understand why these score discrepancies exist. One method is to compare the frequency distribution results of the model variables between time of model development and the current implementation.

Once the application of the solution is within the comfort level of the data miner, they would then need to create a testing and tracking environment in order to evaluate the impact of this solution within a live business initiative.

As we observed in the previous chapter, the intention in the design of a marketing matrix is to maximize results along with maximization of learning.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_17

Fig. 17.1 Example of validating implementation results

% of List	Minimum Score (development)	Minimum Score (application)
0-10%	0.08	0.04
10-20%	0.07	0.03
20-30%	0.06	0.02
30-40%	0.05	0.01
40-50%	0.04	0.004
etc.		

In this case, key pieces of learning would be to test how well the model performs and to determine the effectiveness of various communication pieces or offers that may be tested. As stated in the previous chapter, one cell will represent the winning cell. This would contain the impact of the model, and the expected winning offer, as well as the expected winning communication piece. Implementation is a detailed process. It is important to devote time and effort into ensuring that the solution is correct and makes sense in today's information environment. Additionally, this time commitment also needs to be allotted in order to set up the proper testing and tracking environment for evaluating performance.

Key Takeaway

- Analyzing implementation results enables data scientists to determine whether or not the data environment has changed significantly between time of development and implementation.

CHAPTER 18

Value-Based Segmentation and the Use of CHAID

The process of measuring and evaluating results can also be used to develop an overall segmentation strategy. The most frequently used approach in developing a segmentation strategy is one that is based on value. The key and ultimately the largest challenge in this type of exercise is the identification of those metrics and measures which will comprise value. Once these are determined, a value result is built for each customer record. The results of this exercise are presented in a decile report such as the in Fig. 18.1.

In this decile report, customers are ranked in descending order by value and placed into ten deciles. It is at this point that some judgment is used in deciding the value segment breaks, but this judgment is based on their % contribution to overall value. In this report, the top 20% (the high-value group) contributes 44% of the total value to the entire customer base. The medium-value group (20%–60%) contributes 43% of the total value within the customer base, while the bottom 60%–100% (the low-value group) contributes only 14% of the total value within the customer base. In this example, the customer base has been stratified into three value segments.

It is this placement of customers into deciles that helps us to create the three value segments of high, medium, and low based on each decile's contribution to overall value for the entire customer base. Because these value segments are heterogenous rather than homogenous, the creation of profiles for each value segment becomes meaningless. We could, however, create a profile of what higher-value customers look like.

The other dimension to utilize alongside value segments is one based on behavior. Besides looking at the customer's worth or value to the company, it is also important to consider the recent behavioral trends of the customer which would be reflective of a change in that customer's behavior. This type of perspective allows the marketer to design a program based not only on the customer's worth or value to the company but also on one which attempts to

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_18

% of Customers Ranked by Value	Value Segment	Average Value	% of Value in Interval
0-10%	High	2400	24%
10%-20%	High	1900	19%
20%-30%	Medium	1300	13%
30%-40%	Medium	1100	11%
40%-50%	Medium	1000	10%
50%-60%	Medium	900	9%
60%-70%	Low	600	6%
70%-80%	Low	400	4%
80%-90%	Low	300	3%
90%-100%	Low	100	1%

Fig. 18.1 Value segmentation decile report

VALUE	DECLINE	DEFECTOR	GROWER	INACTIVE	REACTIVATION	STABLE
High						
Medium						
Low						

Fig. 18.2 Value–behavior-based segmentation matrix

deal with the trigger points of this customer's changed behavior. Figure 18.2 lists below is a matrix which allows the marketing decision-maker to select names based on value as well as behavior (a process known as Value-Behaviour Based Segmentation).

Using Key Variables as Segments

Often enough, it may make sense not to model the entire universe for a given marketing program. In previous chapters, we observed that sometimes certain variables are overwhelmingly powerful. Results would indicate that segments should be identified upfront with models being built for those segments.

Correlation analysis, and a quick stepwise on key variables, would reveal that one variable is clearly overwhelming when compared to other variables.

In our example below, correlation and stepwise results reveal that tenure is clearly overwhelming in its impact on the objective function which is response in this case (Fig. 18.3).

Note that the above results indicate that the tenure correlation coefficient, as well as its % contribution to an overall equation, is drastically superior to other variables that are being considered as potential model variables. Chi Square Activation Interaction Detection (CHAID) would then be used to determine the actual ranges of tenure. These tenure ranges would then be used as segments and models would then be developed against each segment (Fig. 18.4).

Variable	Correlation Coefficient	Confidence Interval
Tenure	0.5	99%
Total Spend	0.2	99%
Income	-0.19	99%
Credit Score	0.18	99%

Variable	% contribution to Model
Tenure	90%
Total Spend	5%
Income	3%
Credit Score	2%

Fig. 18.3 Using correlation and stepwise to determine variables for segmentation in optimizing response

	Segment 1	Segment 2	Segment 3
Response Rate	2.5%	8%	4%
# of Names	50,000	300,000	500,000
Tenure	<1	1-3 yrs	3 yrs+
Model	1	2	3

Fig. 18.4 Example of CHAID used to define key customer segments

Here the use of CHAID to define tenure as a key segmentation variable which is further supported by both correlation and regression analysis provides another perspective on segmentation. This is the case where one variable is overwhelmingly powerful in attempting to predict the desired behavior. Segmentation by this variable upfront with customized models for each segment represents the optimal approach here.

In contrast to cluster analysis, where segments are defined homogenously and in an unsupervised learning fashion, this type of segmentation adopts a supervised learning approach with response rate essentially supervising the creation of these segments.

The reason that decision-tree segmentation warrants much discussion onto itself is that it represents a tool with great benefits within the data mining world. First, before discussing the merits of the decision tree, let's understand what decision trees are all about.

Most of these decision-tree tools employ either CHAID (Chi-Square Automatic Interaction Detection Analysis) or CART (Classification Analysis Regression Tree). Both tools again employ the fundamental concept underlying all statistics which is trying to uncover characteristics which vary significantly from a given average. The main difference between the two tools is that the desired behavior or objective function behind CHAID is a binary type of function (0 or 1) such as response while in the case of CART, the objective function is continuous, such as spending or income.

Decision trees are a tool where the actual output resembles a tree-like diagram. Figure 18.5 gives an example from a financial institution regarding an Registered Retirement Savings Product (RRSP) direct marketing campaign.

In this example, the top node (or branch) represents the entire sample or population that is being analyzed. In decision-tree tools, the analysis accomplishes two main deliverables. First, it indicates the key characteristics or variables which are statistically significant against the desired behavior or response,

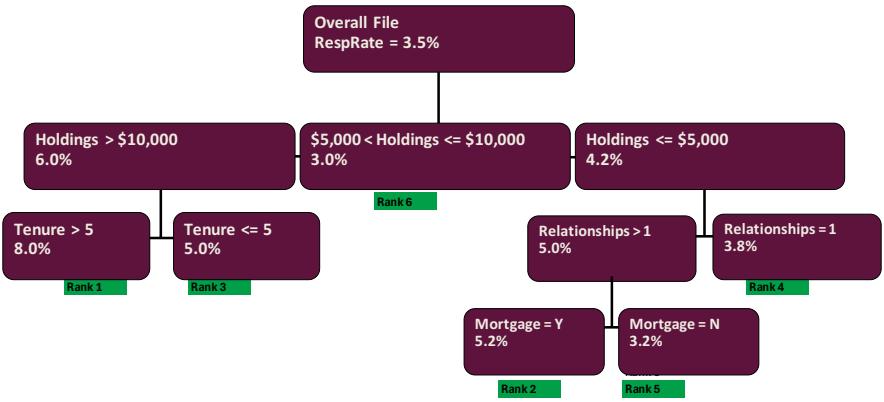


Fig. 18.5 Example of CHAID as targeting tool

as observed in this example. The statistically significant variables represent the branches. The second main deliverable is that the routine determines the optimum breaks or nodes within the variable or branch. Note that the branches and nodes above are determined through Chi-Square routines. The basic math behind Chi-Square is $\chi^2 = \sum(O_i - E_i)2/E_i$, where $i = 1$ to n. Basically, the math is trying to determine the relative difference between what is observed versus what is expected. As this difference becomes larger, the results become more statistically significant.

In our example above, the objective is to optimize response to an RRSP mailing where the desired or targeted behavior is binary (response). The first branch represents holdings and is the most statistically significant characteristic against response. The nodes of this first branch of holdings are, respectively, >10 K holdings, between 5 K and 10 K holdings, and holdings <= 5 K which are determined from the CHAID routine. Once again, the algorithm or statistical routine determines these breaks or nodes for the user. The CHAID analysis, being an iterative-type routine, then continues to look for the next significant variable within each node. In our example above, the node containing holdings >10 k reveals that the next most significant node is tenure, with the following breaks or nodes less than or equal to 5 years' tenure and over 5 years' tenure. Meanwhile the node containing holdings < = 5 k reveals that the next significant node is relationships with breaks of 1 relationship and more than 1 relationship. Meanwhile, the holdings node with 5 k–10 k contains no subsequent node as there is no other variable which is statistically significant from a Chi-Square perspective at this particular branch of the analysis. The analysis continues its iteration where only one more branch is identified. In this case, the node representing relationships > 1 has a subsequent branch related to whether or not the individual has a mortgage. The decision tree is complete at this point, based on the statistical thresholds set by the user.

These kinds of routines allow tremendous flexibility to the user in terms of whether the tree is going to be an exhaustive one with many nodes and branches, or just a simple one with few branches and nodes. Through the use of quasi-"confidence intervals", such as Bonferroni adjustments and sample size minimum volumes, the user can adjust the tree to their own preference. Using a restrictive or large minimum sample size or increasing the cutoff threshold of the Bonferroni adjustment results in fewer variables that are likely to be significant. This creates a tree which is simple with fewer branches and nodes. A more exhaustive type of tree would be produced by relaxing any of these assumptions regarding sample size or the statistical threshold.

Both routines (CHAID and CART) can be used as targeting routines, much like the more traditional multiple regression and logistic regression routines. As targeted routines, the actual end nodes or segments are rank-ordered from top to bottom based on how these end nodes or segments perform against the desired behavior which, in our example above, is response. In our example, we have six segments or end nodes. A gains table or decile table can be produced using these six segments in which the records are ranked by the segments' response rate performance. In our example above, the segments or end nodes would be ranked as follows:

- Rank1/Segment 1: Holdings > $10 K and > 5 years' tenure
- Rank 2/Segment 2: Holdings <= $5 K and with relationships > 1 and who have a mortgage
- Rank 3/Segment 3: Holdings > $ 10K and <= 5 years tenure
- Rank 4/Segment 4: <=$ 5 K and with relationships = 1
- Rank 5/Segment 5: Holdings <= $5 K and with relationships > 1 and who do not have a mortgage
- Rank 6/Segment 6: Holdings between $5 K and $10 K

In our example here, the use of six ranks to create 10 decile groups will cause many records to have the same output result. As a result, this loss of granularity in being able to differentiate records on response rate performance leads to decisions which are less than optimal in trying to achieve the desired business objective. The more traditional regression techniques do not encounter this phenomenon as the output is scores, where much fewer records are likely to have the same score. One option is obtaining more granular-type results from this decision tree is to relax the statistical thresholds, thereby obtaining more nodes and ranks. The negative implication is that these increased end nodes are less statistically reliable. But the user must be able to balance the loss of statistical reliability versus increased granularity. As we have discussed throughout this book, domain business knowledge provides additional insight to help the user make a more informed balanced decision.

Key Takeaways

- Other forms of segmentation beyond the unsupervised approach of clustering exist.
- Value–behavior-based segmentation is a supervised approach that segments names based on value and change in value.
- Decision-tree techniques such as CHAID represent a modeling approach where the output is segments rather than individual scores.
- Segmentation can be used in conjunction with predictive models where customized models can be developed for each segment.

CHAPTER 19

"Black Box" Analytics

Organizations are now aware of the significant contribution that predictive analytics brings to the bottom line. Solutions, once viewed with skepticism and doubts, are now embraced by many organizations. Yet, from a practitioner's perspective, certain solutions are easier to understand than others. The challenge for the practitioner is to provide a level of communication to the business end-user such that the end-user has the requisite degree of knowledge when using the solution. What does this mean? From the business end-user's standpoint, this means two things:

1. The ability to use the solution to attain its intended business benefits
2. Understanding business inputs

Under the first point, of being to attain the business benefits of the solution, the use of decile reports/gains charts and AUC curves have been previously discussed as they do provide the necessary information to make the correct decisions when using these solutions. The ability to rank-order records with a given solution yields tremendous flexibility in understanding the profit implications under certain business scenarios.

The second point of understanding the business inputs can be a challenge as it is here that the analyst attempts to "get under the hood of a solution" and actually demonstrate the key business inputs in the model. Yet, as the world of predictive analytics has evolved, so have the techniques. Highly advanced mathematical techniques using approaches related to artificial intelligence (to be discussed later in this book) present real challenges in trying to explain the key components and triggers of a model. What do we mean by this? For techniques such as multiple regression and logistic regression, the solutions are linear and sigmoid, respectively. For techniques such as artificial intelligence (AI) and its basic neural net mathematics, the solutions are often non-linear. For example, suppose there is a distribution where a variable like tenure has a

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_19

trimodal distribution (that is, there are 3 mode points within the distribution) with response. How can one meaningfully explain the trend to the business user? Yet, in many of these advanced techniques, there are several variables which can exhibit this high degree of non-linear complexity. With some of these advanced techniques and software, equations are not even the solution output. Instead, the practitioner is presented with output in the form of business rules. Further complicating this scenario is the fact that many of these non-linear variables may have interactions with each other whereby the actual interaction between multiple non-linear variables represents an actual input to the solution. These types of complex inputs pose real difficulties to the practitioner if they are trying to educate the end-users on what the model or solution is comprised of. The practitioner's typical response is that the solution is a **"BLACK BOX".**

Yet, besides the communication challenge of "black box" solutions, the second challenge is credibility. Will the business community really trust that the equations and variables as seen in "black box" solutions are simply superior in explaining variation than in the more traditional linear and log-linear techniques? Mathematically, it may be possible to explain how a given input best explains the variation. But can it be explained in a business sense? For example, if there is a linear relationship between tenure and response, it is far easier to explain that longerhigher-tenured people are more likely to respond thereby explaining why tenure has a positive coefficient within a given model equation. Conversely, assuming a linear relationship between income and response, it may be concluded that higher-income people are less likely to respond, which thereby explains the negative coefficient of income within the overall model equation. But the ability to explain the impact of tenure on response becomes more difficult if the trend is a curvilinear one with multiple modes within the distribution. Tenure in this case would be captured in some kind of complex polynomial function that is very difficult to explain in business terms.

With these types of "black box" solutions, user-driven parameters provide the necessary flexibility to the practitioner in delivering a variety of different solutions. Without any test or validation dataset, the practitioner can alter the parameters in such a way as to deliver an almost "perfect" solution since these high-end mathematical solutions will attempt to explain all the variation, including even variation which is truly random. This is certainly the case with neural net models.

Of course, the key in really assessing the superiority of one solution over another is validation, and the resulting gains charts where one can observe how well one solution predicts behavior over another solution. The notion of a robust validation environment is one area in which practitioners and academics are in complete agreement.

Yet, this inability to explain the key business inputs can present a barrier to broader acceptance of using these tools throughout the organization. Businesses need to have a "deeper" understanding because of the measurement process

used to evaluate the performance of a given solution. This process cannot occur effectively if there is a gap in understanding what went into the solution. Of course, this is not a problem if the solution works perfectly. But what happens when solutions do not work? Under these scenarios, deeper forensics is required to understand what worked and what did not work, which certainly implies a deeper understanding of the key business inputs. This comprehension gap in being able to effectively measure "black box" solutions is the reason why most organizations opt for simpler linear or logistic-type solutions. In this type of environment, the notion of "less is more" is very applicable since these non-"black box" solutions are more easily understood and, most importantly, can be analyzed in a very detailed manner, particularly when these solutions fail to produce the expected results.

Recent advancements, though, in data science have explored the issue of feature importance in a given solution. This is particularly relevant with solutions that are now more AI-driven. Solutions now look at the change of variation between predicted and observed output with the presence of a given feature versus the non-presence of the given feature. In this approach, the feature or variable looks at all its permutations and its specific change within a given overall solution. It then sums up all these changes for that variable versus all the total changes for all the other variables in the dataset to arrive at the importance of the feature. One of the more common methods is the Shapley approach which mathematically looks at the marginal contribution of each variable relative to other variables within a given solution. These approaches work quite well, particularly in the area of image recognition as they can demonstrate or highlight the features within an image that are most important in the solution.

Although much work has been recently done in this area, mathematical mechanics is not too dissimilar from the charts that were displayed every time we developed a model using either multiple regression or logistic regression. (See Chap. 12, Fig. 12.11.) The math here used the partial R^2 as a means of calculating the % contribution by each variable within the model or its feature importance. Obviously, the complexity of neural net models have posed challenges in this area which were non-existent in both logistic regression and multiple regression models. But the researchers, as discussed above, have devised approaches that have, for the most part, overcome these challenges. Through this recent work, "black box" solutions are simply not an acceptable outcome, no matter the complexity of the solution.

Key Takeaways

- "Black box" solutions, particularly in the more advanced neural net models, have historically been an issue.
- Recent research has attempted to focus on how much of the difference between the observed output and predicted output is caused by a particular feature.

- This above research has mathematical similarities to what we have seen when using partial R^2 in trying to explain the % contribution of a variable within a logistic or multiple regression.
- The Shapley technique is a very popular approach and one that has been used extensively to determine the importance of features within an image as well as consumer behavior models.

CHAPTER 20

Digital Analytics: A Data Scientist's Perspective

The world of marketing has undergone a fundamental shift in the last 30 years with the advent of digital marketing and social media as new forums within which to interact with customers. Digital technology now allows marketers to communicate in a variety of different ways. Besides being just reactive to communication, digital technology has enhanced the consumer's ability to be more proactive in terms of both what they desire and how they expect to be communicated with. The consumer now has much more control over the marketing landscape. Companies that fail to recognize this new paradigm do so at their own risk. Email, programmatic advertising, social media communication, mobile texting, GPS alerts, and blogging represent just some of the many different facets of communication that were not in existence in the 1990s. Additionally, this does not include how a consumer interacts with a given company's website.

The one commonality among all these different facets is that communication is one to one which, for many pre-internet direct marketers, represents familiar territory. The ability to use information and to act on it has always been the overarching philosophy of direct marketers. From this philosophy, direct marketers have identified three pillars or principles which are the keys to success in direct marketing, but also to success in digital marketing. These three principles are:

- List (What is the target audience of my communication?)
- Offer (What is the product or service that will fulfill a need to the end consumer?)
- Message (What should be the content of my communication to the end consumer?)

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_20

Of the three principles, the most important is the list or target audience in achieving success in both direct and digital marketing. In fact, analytics and data science efforts are primarily focused on this area, as seen through the growth in predictive modeling and advanced forms of segmentation such as clustering. Within the digital sphere, however, let's first talk about email.

Analytics Within the Email World

Capturing information concerning email is like the direct mail medium, albeit that it is much more automated. Presuming, for example, that we had conducted an exercise to group emails into cell code categories, marketers can track who received the email, when it was sent, and the type of email. On the response side, response activity can be measured based on who opened the email (open rate) and who responded to a call to action, such as accessing a URL embedded in the email. Responses to this call to action would be measured by the click-through rate. Besides the increased automation of data capture within email, the other advantage is that information is instantaneous since, in theory, all email recipients simultaneously receive the email upon its deployment. This instantaneous ability to access information on the internet has resulted in new terms such as the ability to conduct tasks in real time. From an analytics perspective, this implies capabilities such as real-time reporting and real-time analysis. Once the analytical or reporting objectives have been defined, there is no longer any bottleneck or time lag in terms of waiting for the data to do any analysis.

Given the real-time nature of information on the web, and its importance to businesses in terms of more effective decision-making, it is only natural that the early pioneers within the web development space would attempt to provide services in being able to properly track online customer behavior. The early adopters (or pioneers) developed reports that allowed companies to track such key online metrics as open rates, click-through rates, bounce-back rates, and conversion rates. From an email standpoint, success was immediately determined by the ability of a given organization to maximize open rates and click-through rates, as well as conversion rates, while minimizing bounce-back rates. One thing to remember in relation to conversion rates, though, is that they may not always be accurate, particularly if a customer chooses to buy through a different channel. For example, I receive the email and decide to purchase the product at the store location.

Although the above-mentioned metrics do provide an overall glimpse of whether or not the campaign is successful, the ability to conduct deeper analytics is somewhat limited. Experienced direct marketers are always interested in customer online behavior prior to the campaign, such as the recency of the last email sent, the frequency of emails and, of course, type of emails. More importantly, we are interested in observing how this prior online email behavior of customers will impact their behavior within the current campaign. Yet, for many organizations this kind of information needs to be captured in some kind

of a marketing database which, once again, is very familiar to experienced direct marketers, but not necessarily to specialists within the online world. Yet, it is this rich historical information of prior promotion activity which will really yield valuable learning concerning the online trends and behaviors of customers and how it impacts future email behavior.

Analyzing Web Browsing Activity

Having provided some discussion about how analytics might be considered within the email world, it is also important to look at the reactive side of online marketing, such as individuals browsing on a company's website. Browsing a company's website is very similar to viewing a TV or billboard ad, except that this browsing behavior is readily available as information to the given company. But the first issue is how to gather this information effectively. This implies that we have the ability to effectively tag customers as being unique when they arrive at a given website.

In the early days of the web, tags were created that were based on the Internet Protocol (or IP) address of the computer. This tag supposedly represented the unique identifier of a given customer. There were, however, flaws or complications within this type of system which mitigated one's ability to identify real unique customers. For example, multiple computers could have the same IP address, unique customers could also be using multiple computers with different IP addresses, and IP addresses might be dynamic, which implied that these addresses were changing at certain time intervals as determined by the ISP provider. As a result, the information collected from these devices was inconsistent.

The most common approach, though, in collecting digital information about individuals is through the use of so-called "cookies". Essentially, an individual clicks onto a given website, which then assigns a piece of unique information, called a cookie, to the browser of that person's device. The limitation here is that individuals typically have multiple devices. For example, I have both a laptop and a mobile phone and my browsing behavior is very different on these devices, meaning that I will receive different types of ads on each device.

The best option in gathering digital information about the individual is to utilize a user authentication-type system which asks for the person's email and perhaps some other demographic information. The customer will also be asked to put in some type of password at the first point of contact. Subsequent visits to the site will simply ask for the email and this password, which then maps this visit to the correct unique customer rather than to the device. This information can then be overlaid on the customer database to capture the customer's online behavior.

With a better ability to recognize unique visitors to a given site, the next issue is to really understand this type of information and, specifically, the source information. The source information of visitors to a website is contained in log files, which contain the following information:

- User ID
- Time of login
- Click/page request
- Status code, which refers to the outcome of the request
- Size of the object referred to the user
- Referrer (where the user came from)
- User agent browser

The status code and the user agent are probably the least important pieces of useful information in any web analytics type exercise. For those of us who have a large amount of history in working with data in the offline world, the above log files appear to be somewhat like transaction or billing files which are seen in traditional database legacy systems. Recency, Frequency, Monetary (RFM), as well as transaction-type variables which are created in typical transaction files, can be created from log files and represented as:

- Recency of last login
- Number of logins within a certain time period.
- Duration of session
- Types of page view
- Page view behavior over periods of time.

Again, the limitation for many current providers in the web analytics space is the inability to provide an historical perspective in prior campaigns that have been sent to the consumer. Digital marketers can certainly assess the prior browsing behavior of the customer. But they cannot determine the impact of prior emails, prior text messages, or other types of digital ads on my future browsing behavior. As mentioned in the email example, a marketing database (or data mart) is the required tool which can provide this capability. Experienced marketers have long become accustomed to marketing databases that provide this ability to assess the sensitivity of prior campaigns on future campaign behavior. The development of this type of information capability has been often referred to as "Campaign Management". Detailed digital information is now accessible to determine the granular browsing behavior of a given customer, but lacks information on what a given consumer has been sent from a given company.

Log file data provides rich information concerning page view behavior as well as recency and frequency in terms of visit to a given site. Additional pieces of information provided by web data which are unique to this environment include the length of time that a visitor spends on a given site. With all this rich information, models can be built to predict the propensity of any click behavior with clicking behavior being a page view, an order, a response to a survey, or perhaps a link to another site. Since time of engagement on the web is captured on the web, both pre- and post- windows can be created in the analytical file when attempting to build predictive models. The post- window, as previously

discussed, would contain either the behavior to be predicted or the objective function such as a page view, URL, etc. Similarly, the pre- window would look at all web behavior prior to the post- window. Given the quantity of data which is available, the challenge is to identify and create meaningful variables from this very rich information. Once again, as seen in the offline world, defining the relevant universe for modeling is a critical preliminary step before building any model. Preliminary characteristics that filter out visitors would be used to identify customers that are clearly not relevant for the modeling exercise since this group are very weak performers in terms of the desired modeling behavior. Since online models can be scored in real time, these types of visitors would simply have no score, implying that they represent the worst names. Model scores can be translated to ranks, thus indicating to the end-user the relative degree of model performance when compared to the average.

Given the speed of the access to digital information, expectations on insights from the data increase. Yet, one can easily become overwhelmed by all the data sources from the web:

- Mobile
- Website
- Social media posts and tweets

The challenge for the data scientist is to link all this data to the individual level. Given this challenge, more robustness in terms of standardizing the "analytical environment" is required. With a standardized data environment, analysts can spend more time creating tools and conducting analysis that allow organizations to enhance their ability in gleaning business intelligence from the data. The data environment under these conditions yields the capacity for developing at least a hundred or so models within a given year simply because there is a standardized analytical file. All potential inputs to any model would be identical for every exercise. The only difference under each modeling scenario would be that the objective function or desired modeled behavior would change. The largest constraint here would be the labor resources to build these tools along with the measurement and tracking of these tools.

Having reviewed the different forms of interaction within the digital world, the real issue, as with the offline world, is whether the information is available at the individual level. The digital collection of data has provided us with an intensive data environment where individual-level data is more readily available. Advanced mathematical tools, particularly the use of AI, are now becoming common business practices through the data-rich environment of the web.

Big Data

The discussions and debate surrounding Big Data, which has been another outcome of the internet and its development, present new challenges to the data science community. In the Big Data vernacular, the three V's (volume,

variety and velocity) continue to emerge as the key challenges in both processing and analyzing the data. For experienced marketers, large volumes have always been the typical environment when attempting to develop and implement successful marketing programs. However, the notions of variety and velocity of data represent the newer marketing challenges. On the variety front, data now arrives in semi-structured formats whereas most experienced practitioners are extremely well-versed in traditional database marketing technology of rows and columns. Yet, before I discuss the newer facets of Big Data, a short discussion on database marketing technology is warranted, especially since this expertise and knowledge will still be very applicable in the world of Big Data. Typically, data arrives in the following fashion (Fig. 20.1).

In the above example, we have a customer table and a transaction table. The structured format is determined by the fact that the records (three for the customer table and three for the transaction table) represent rows. Meanwhile, the characteristics of the record or variables represent columns. Typically, there would be a way to link these two tables and, in fact, the customer number on the transaction table here (not shown here for simplicity purposes) would be the match key in linking these two tables.

On the internet, much of the data is semi-structured or object-oriented. Here the variables are not columns but rather objects. Instead, we have blocks of data which are our raw sources of data to be analyzed (see Fig. 20.2).

In the example above, this block of data represents one user tweet. All the variables in this data are objects. I have highlighted five of them. Four of them highlighted in lighter shade are SOURCE, ID, LANG, and CREATEDATE. The fifth variable highlighted in darker shade represents the actual content of the tweet which is listed as TEXT. The content in the TEXT field is what we call unstructured data.

The semi-structured nature of this data yields records that will have varying formats. This varying formats on twitter data imply that objects or variables will not necessarily appear on each record. Adding to this complexity are the many other social media digital platforms as well as the different mobile devices that gather data about the user.

The varying formats of data have created the need for technical capabilities in extracting information from the web. To address this need, social media vendors provide Application Programming Interface (APIs) to help read the

Structured Data

Customer Nº	Household Size	Postal Code	Income
1	3	L1A3V1	125000
2	2	M5S2G1	30000
3	1	H4B2E5	40000

Transaction Nº	Date	Amount	Product Type
1	JUL 15-2009	100	A
2	OCT 1-2009	75	A
3	SEP15-2009	200	C

Fig. 20.1 Example of structured data

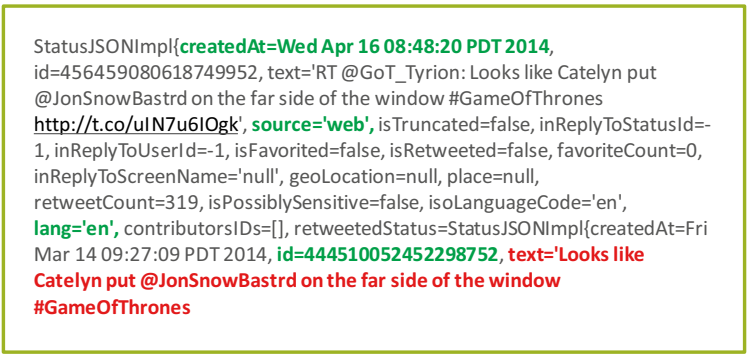


Fig. 20.2 Example of semi-structured Twitter record

data that is arising in these constantly changing file formats. For those users that are more technical, programs such as Python, Java etc. allow the analyst to extract this type of information without using any software vendor tool.

One other additional complexity, the 3rd 'V; of Big Data, is velocity. This addresses the notion that much of this data in the digital world is constantly streaming. In other words, the data is constantly "coming at us". From another perspective, this velocity of data simply accelerates the volume of data that needs to be processed.

As we observed that new Extract Transform Load (ETL) tools are required to handle data in ever-changing semi-structured file formats, new tools are required to handle the "data processing" requirements in order to deal with both the volume and the velocity of data. Traditional sequential data processing routines are no longer adequate. Techniques involving parallel distributed file processing and Map-Reduce-type technologies have allowed organizations such as Apache and its Hadoop technology to become a mainstream name amongst present-day businesses. Rather than get into the technical details of what this really encompasses, its end deliverable is the ability to process data at significantly reduced timeframes than the traditional techniques of file sequential data processing.

Integrating both these new data-processing techniques alongside the ETL techniques enables data science practitioners to operate in the Big Data world. As has been stated throughout this book, however, it is not about data for data's sake but rather data to help solve business challenges. Adopting a "So What?" attitude to data helps one to understand what data is really required for a given business problem. Is every problem a Big Data challenge? The answer is "No" as the real answer is that every problem is a data challenge, be it Small Data or Big Data. In other words: what data do we really need to solve the business problem at hand?

Key Takeaways

- The internet and digital data have been both a blessing and a curse to the data scientist.
- Access to data is almost limitless, but determining what is useful is even more of a challenge.
- Digital data is primarily semi-structured or object-oriented, which has enhanced the use of programming tools such as Python and Java.
- With extremely large volumes of data becoming the norm, parallel-type processing technologies, such as Hadoop and Map-Reduce, represent the best-practice option.
- The data-rich digital environment provides tremendous insight into the user's browsing behavior.
- Yet one gap is the ability to understand the impact of prior campaign history (i.e. what promotions were sent to the visitor) and its impact on future browsing behavior.

CHAPTER 21

Organizational Considerations/People and Software

The world today is experiencing tremendous change in virtually every facet of society: political structures, as seen by the fall of communism but to some extent increasing autocracy; sports with their new collective bargaining agreements; businesses with their obsessions to become flatter and leaner, all with the intention of reducing head count; and technology in its all-encompassing goal of making us more effective with our use of information. Of these, the trend that is most clearly demonstrated within the world of technology, is data science, which strives to embrace the use of the best and most relevant technologies to solve a business problem.

Among the areas which have changed over the years, and which will continue to change perhaps in a more transformative manner due to AI, include:

- Organizational change
- People
- Software

With data and information becoming an ever-increasing corporate asset, traditional approaches to corporate structure are becoming less and less relevant. Technology and increased knowledge empowerment are flattening the organization and, ultimately, reducing the number of layers between the CEO and the average employee.

These types of changes are not merely isolated to corporate structure, they are also critical in creating new roles and responsibilities. Recognizing that knowledge and intelligence, and its use for better decision-making, is the real end deliverable behind data and information, there is a growing need for a chief knowledge/chief analytics officer. This role is not necessarily an IT role, as such roles typically embrace a deep understanding of technology and systems with the end deliverable being access to data and information. Using data to

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_21

derive meaningful knowledge and insights requires a different type of skillset. The ability to think about data and information in this fashion is vastly different to that which is required to think about how data is best disseminated to users. Of course, the knowledge or data analytics officer needs to work closely with the chief information officer (CIO), but both roles are equally important if information is going to be the key competitive asset of a given organization. What, then, should the chief knowledge officer or chief analytics officer look like? One option adopts a more centralized approach where all the analytical areas report directly to the chief analytics officer. For example, in a credit card company, all the marketing analyst areas as well as the credit risk/fraud analysis areas would report to this person. Although the analytical objectives of each area (marketing and credit) would be very different, the one common theme is that both functional areas use data to derive insights for better decision-making. One of the primary roles of the chief analytics officer (CAO) would be to somehow synergize the analytical activities of both areas. This means that analysis should be focused on optimizing marketing behavior while at the same time reducing credit card/fraud risk. The performance objectives of each functional area would, of course, be prioritized on how they perform within the area. But a significant portion of their performance evaluation would also be based on the performance of how the other areas are performing. Through this cross-fertilization of goals and objectives between functional areas, the chief knowledge or data officer is able to better integrate the activities and tasks between areas so that they are aligned with the overall corporate goals.

Data Science in Marketing

Historically speaking, no departments essentially existed dedicated to data science. As direct marketing evolved into a more common marketing discipline, marketing service departments were created to help in the execution of direct mail programs. Specifically, this involved several activities. The first activity was to generate names for a given campaign by working with list broker specialists who would recommend specific lists based on the desired campaign objective. The second activity involved data hygiene and cleansing to ensure that the name and address were correct, and that there were no duplicate names on the promotable file. The final activity involved the actual generation of the campaign list file and its accompanying test cells. In addition to targeting the best names for a campaign, the test cells were used to derive specific learning which had been identified upfront as one of the campaign objectives.

From an analytical perspective, minimal activities were conducted within the direct marketing area. As direct marketing began to grow in prominence, however, the analytical component continued to evolve and grow due to increased demand for better information and insights. Ultimately, this necessitated a requirement for stronger mathematical skills and a certain level of statistical knowledge. With the analytical component being able to deliver and demonstrate significant tangible benefits directly to the bottom line, there was an

increase in both the volume and the complexity of work. This growing demand ultimately resulted in the creation of either data science or CRM analytical departments. Typically, in many organizations today, the analytical areas report separately to the functional areas. Examples of this include the marketing analysis and credit risk analysis areas, which operate as separate silos. But as stated earlier, some organizations are centralizing the analytics/data science function. The intention of this centralization is to allow the organization to have a more holistic view of the customer and, ultimately, customer profitability. More of this consolidation of data analysis activities is expected as increasing numbers of organizations look to create customer-level profitability. In some cases, they are adopting a hybrid approach where analysts have a solid line reporting role within the analytics area, but maintain a dotted line reporting role to the respective functional area where their skills are being utilized. However, our experience has revealed that most organizations still adopt a more decentralized approach towards analytics and data science, where there are separate analytical/data science units within the functional areas of an organization and no centralized analytics department.

Given its tremendous volume of data, the internet will also reinforce this trend in data analytics consolidation. The key will be to somehow merge the online and the offline world of data and, more importantly, to determine how to use this merged information for better decision-making. The world of Big Data and digital analytics will simply augment the need for a more centralized perspective on analytics and the requirement for the CAO. Even without this centralization, the new Big Data paradigm will enhance the significance of data science and its role within the organization.

Outsourcing to Present-Day Industry

The notion of outsourcing vs. insourcing various company tasks continues to be a sensitive topic throughout society. The academic, business, and government communities have all contributed different perspectives on this topic. With increasing technology and globalization, it is simply easier to outsource functions now in ways which were impossible to do ten to twenty years ago. One example of this is the telemarketing functions performed on behalf of many large organizations. Increasing globalization now allows organizations to identify not only where tasks can be achieved most inexpensively, but also where the revenues from these tasks can be maximized. Regarding telemarketing activities, many organizations have determined that it is cheaper to conduct these activities in certain countries with lower labor costs. The costs of training are also reduced as the education level within these countries continues to improve. Meanwhile technology has allowed communication to be almost seamless in the sense that one does not know if the phone call originated in Toronto or New Delhi. These kinds of conditions allow organizations to maximize their profit by selling to the North American market, but marketing it via a low-cost country.

Another example of this is the technology sector itself. Traditionally, technology solutions have always cost money. The cost of a programmer without experience typically ranges from a minimum of $35.00 per hour (including benefits) and perhaps going to be well over $100, depending on the breadth and depth of experience. Once again, low-cost countries in Asia purport to offer these skills at well below the going North American rate. For many companies, certain technology solutions are farmed out to these low-cost providers. However, what is interesting here is that not all solutions have been farmed out. At present, it is still the case that those technology solutions which require heavy intellectual capital are conducted in North America, while the more "routine" solutions are developed overseas. An example of this might be the design of a new system to meet the needs of the business users within a given organization. The research and design (R&D) would be done in the North American market while the actual coding of screen and visual interfaces might be done overseas.

The advent of generative AI (GENAI) tools is also revolutionizing this area as the programming code can automatically be generated in a variety of languages through the use of the correct prompts. The critical knowledge component here is the generation of the right prompts to write the right programming code.

Evolution of Outsourced Data Science Analytics

The data science industry is certainly not immune to the debate concerning outsourced activities. Like many other services, the need for data science developed internally within certain organizations that felt it was a critical component to their business. Even going back to the early 2000s, there were only a handful of companies which were attempting to do this kind of work. These few organizations invested heavily in the technology and resources that were required at that time to engage in data science activities. Many people working for these types of organizations recognized the tremendous business value of data science, but realized that technology was indeed the limiting barrier at that time. This all changed in the 2000s as technology barriers were eliminated, thereby allowing many more organizations to conduct data science activities as part of their core service deliverables.

As organizations are more empowered in the data science area, they need to determine if activities should be outsourced in full, in part, or not at all. The solution reached will depend on the type of organization but also, more importantly, on the individuals that are the key stakeholder decision-makers.

Core Versus Non-Core Activities

At an organizational level, company activities and tasks are segmented into core and non-core competencies. For companies that consider data science to be a non-core activity, the solution is simple. In almost all cases, much of the work

will be outsourced while being managed internally by the company. Some of the more mundane data science activities, such as basic-level reports, may be produced externally, depending on the data and analytical skills of the external personnel.

Meanwhile, companies that deem data science to be a core competency will be more reluctant to outsource either the mundane or the more advanced activities. These core competencies are often considered to be strategic in nature, and therefore a critical competitive advantage of the company. In determining the core nature of data science as a business activity, the following questions could be asked:

- Does data science provide critical intelligence and business advantage in the selling of a company's given products/services?
- Do I need data science to provide the necessary strategic insights which will maintain our organization's competitiveness?

For these types of organizations, the answer is "yes". If data science is a core competency, organizations will seek to bring all their capabilities in-house. External data science consultants would then be hired to provide a different point of view or to conduct some specific act or activity where the knowledge competency is lacking.

Challenging Internalization of Core Data Science Activities

For today's larger institutions, no one will refute the strategic importance and, ultimately, the core nature of data science as an overall business activity. Yet, it needs to be understood that its many deliverables are often tactical in nature. For instance, developing models to target names is a tactical solution to a specific business need. Certainly, the importance of models as being a very critical component in the provision of services and products to an organization's customers is a business reality. But does this type of work always need to be done internally? In other words, does the company lose an advantage by outsourcing this function to another company? Certainly, the outsourced company will have information pertaining to customer characteristics that are most pertinent for a given customer behavioral model. But this information is unique and specific to that organization, implying that this learning is not easily transferable to other organizations. This is the real crux of the issue. If the learning from a data science exercise is really nontransferable, then it does not matter if solutions are developed internally or externally. However, does the data science exercise provide any solutions which can be transferable to other organizations?

At a very broad level, one could argue that key customer groups might be transferable to other organizations within the same sector. For example, it may

be determined that mortgage buyers represent a distinct customer segment within a certain data science exercise for a financial institution. However, even using this same definition of segments would be suspect as it runs counter to the basic principles of data science, which is to allow the organization's own unique data to define the segments. Furthermore, the segment definitions would use unique business rules that are customizable to the company's own data.

The solutions themselves are not transferable in nature between organizations. About the data scientist's knowledge, should this be internalized or relied upon externally? One perspective on this states that the company best utilizes the knowledge base of data science for its specific business needs. This would imply that the goal of internalizing the knowledge base, as opposed to using external resources, is not the real issue. The important considerations here are where to obtain the best value and services, and how to do it inexpensively. From an organizational standpoint, executives should ask the following questions when deciding to outsource data science:

1. Can I meet all the demands for this work internally and at lower cost?
2. Do I have the required expertise to do this work optimally?
3. Does a different point of view and perspective yield insights that can further optimize business decisions?
4. Does a complementation of in-house resources and outside expertise deliver incremental value?

The discussion to date has attempted to deal with outsourcing purely from a business rationale standpoint. However, there are, at times, more subtle and hidden reasons that can impact the decision to outsource. These reasons are often more psychological, and can be attributed to people's relative reactions to security. All of us desire to be secure in whatever environment we find ourselves. Certainly, job security is one specific example of this. Here is an example of how this psychology might work. Suppose a person, John Smith, runs a marketing services department consisting of ten people. John is told that outsourcing needs to be considered because there is more demand for these services, but that these services need to be completed with fewer internal people. His first reaction is the realization that the value of the department to the organization has been diminished due to these outsourcing considerations. There is also the fact that he is expected to do the work with fewer people. Rather than viewing this issue strictly from a headcount perspective, however, he could view it from a budgeting perspective. In other words, the real significance of his department relates directly to the budget that is accrued to his area, and not necessarily the number of staff. However, he needs to determine how to influence the dollars allocated to his budget. Two things need to be done:

- Demonstrate the benefits of data science activities and its impact within the organization
- Ensure that the senior people within the organization clearly understand these benefits

Headcount increases within the department alone will not accomplish this. Yet, if John is focused on maximizing the benefits of data science within the organization, then he will try and find the best resources, both internal and external, to achieve this. If John can demonstrate benefits to the organization that result in $15,000,000 being allocated to his area, it does not matter whether he can achieve these benefits with 15 employees or 1 employee. The priority is to get it all done within the given timelines and to obtain the best solution using both internal and external people. The adoption of company policies which exclude outsourcing as an option can be very limiting. Even with highly talented internal people, external suppliers, given their broad range of industry experience, can bring a point of view and perspective that yields unique insights into a given business problem.

Certainly, one important component in addressing job security is not to build up an empire in terms of headcount but rather to build up the required budget for this area. Budget increases are essential to job security. But the real key to maintaining budget amounts, or even to obtaining budget increases, is to continually achieve the benefits of data science, which is best accomplished through a mix of both internal and external resources.

Another growing trend in outsourced analytics is the option of conducting advanced analytics offshore. As discussed previously in the telemarketing example, offshore firms now conduct analytics which was traditionally the sole responsibility of the onshore firm. The types of activities tend to be more technical in nature, with heavy programming being required along with the application of advanced mathematical techniques. Meanwhile, the onshore company has the depth and breadth of data science experience across a wide range of industry verticals, thereby allowing it to manage the process more efficiently by not conducting all the requisite executional data science tasks. In practice, however, there have been mixed reactions about the efficiency of this process in terms of streamlining costs for a given data science project. The question which organizations need to answer is whether it makes sense to hire junior data science analysts who initially would execute much of the same activities as the offshore organizations. One might argue that the in-house training and mentoring of these junior analysts might be considered an investment as the knowledge and experience of these junior analysts would be superior to what might be obtained by offshore analysts working with onshore companies. Onshore organizations must never stop investing in new analysts from their own shores as the potential payback is great. The deeper problems that need to be solved through data science can only be resolved by individuals steeped in both a high level of academic training and also practitioner-based expertise achieved though mentoring and job training. Yet, at the same time, offshore

companies do need to be considered as viable options where the tasks, albeit very technical, tend to follow processes that are clear and straightforward.

Once again, the dynamics of GENAI are providing another perspective on the irrelevancy of outsourcing vs. insourcing considerations. For many of the less advanced data science tasks, I can use GENAI as my option instead of some offshore company. And I can also do it quicker, thereby allowing my onshore team to devote more time to the more advanced data science tasks.

People Factor

This has always represented the largest challenge in building a data science practice. Since the data science discipline is relatively new, it has been difficult to find knowledgeable practitioners within this area. In the past, organizations typically relied on finding individuals from leading-edge organizations with extensive data science experience. Formal educational training in this area was historically non-existent.

This has obviously changed as most universities and colleges across Canada now offer specific data science courses, as well as business degrees with an emphasis in data analytics. Several institutions, such as Queen's University, even offer master's programs. In fact, the more recent developments in this area have resulted in several institutions offering courses and programs specifically in AI.

Besides the increased academic focus in this area, industry itself has also provided a variety of forums, as demonstrated by the plethora of seminars and conferences that deal with topics in this area. Educational platforms such as Coursera offer a more customized self-teaching approach to gaining knowledge in this area. At an even more basic level, one can Google as well as use GENAI tools to gain specific detailed knowledge on any topic. In some senses, the practitioner has never been more empowered. The only limitation is the desire and motivation to seek more knowledge in these areas.

All the above education, through academia, industry forums and online educational platforms, will always be inferior to that valuable experience gained "on the job". Only industry experience and being "hands on" with a live business problem and data can yield that unique practitioner knowledge which is so valuable in the marketplace.

Software Changes

The software component of data science remains the area of this field which continues to see the most change. Why is this so? With cheap and easy access to data through technology, there has been huge demand in trying to empower more people to do data analysis. Historically, this type of capability was isolated to highly technical people with strong programming skills. In the very early days of data science, those with SAS programming skills were treated as the data science gurus within any company. These skillsets are still highly valued,

but interest has moved on to other tools such as Python, which are more compatible with the Web. At the same time, there has also been a shift to empower other individuals in the area of data science. The first scenario is to empower regular business analysts with the capability of doing non-statistical data analysis. In the second scenario, organizations would like to empower more mathematically oriented individuals to conduct statistical-type analyses, yet their capabilities are limited because they do not have strong programming skills. In both of the above scenarios, new software can help to facilitate the data science exercise by eliminating the need for someone to write programming code. Instead, the Graphical User Interface (GUI) provides the interface that allows the analyst to conduct a data science exercise. The analyst still needs to have a very deep understanding of the data science process, but they can now conduct this exercise without writing any programming code.

Another significant improvement within the data science software world is the development of tools that purport to conduct real-time analysis. In other words, analyses can be conducted at any time with the most up-to-date accurate information. This can have tremendous relevance within the online world, where information is being exchanged constantly. Having tools that can both access and conduct analysis at any time represents a very significant improvement when compared to the past practice of having to wait for the next mainframe database update, or wait six to eight weeks before having enough data to analyze a direct mail campaign.

Much of this new software now has high-powered mathematical tools to help target specific prospects and/or customers. Although it is exciting as a data scientist to have access to these new products, it must never be forgotten that they are indeed just tools, and that they are no substitute for skills, knowledge, and experience.

Software vendors and high-powered mathematicians might be inclined to say that their new tool is the next breakthrough in mathematics; however, it is important for data scientists to be "grounded" in how they evaluate these tools. Proper validation exercises are critical towards effective evaluation of these tools.

Text Mining and Text Analytics

The development of text mining within data science now represents a process where text data can now be mathematically analyzed. The learning and work being conducted in this area include techniques that focus on processes to convert unstructured data, such as text into actionable structured data. This is done through a process involving data parsing/cleansing, the conversion of the data cleansed/parsed data into actionable numeric data, and then the classification of this numeric data into topics or categories. Consider the opportunity here. Data scientists can now use free-form typed data, written comments, and other text-type data on the customer to uncover certain customer patterns that might be relevant for a given behavior. One example of this is the ability to

examine email data between customers and a retail company. With this data, text mining analysis may indicate that classification should be into three categories: complaints, request for catalog information, or request for specific product information. This could then be used to determine if any of these three categories have an impact on retention. The ability to identify and discover patterns in this type of text data is a growing area, as the number of social media platforms and their respective tweets/posts continues to expand within the digital world.

As the data science industry grows, products continue to be developed that empower more business users within this discipline. These tools enable both high-end business users (statisticians, programmers, mathematicians) and low-end business users (business analysts) to conduct data science tasks more efficiently. For the high-end users, sophisticated statistical analyses can be employed very quickly. They can try several different mathematical approaches and observe the results. The software in this case is allowing the analyst to spend more time on analytical thought. This kind of thinking would allow the analyst to use his or her interpretative skills in evaluating results across several different approaches. In such exercises, five different techniques might be utilized: logistic regression, CHAID, neural nets, genetic algorithms, and linear regression.

At the same time, suppose a data environment is to be created that does the following:

- create index variables on all continuous variables
- create change variables on all variables which have both a time and value component attached to the variables
- create categorical variables using CHAID
- create categorical variables using factor analysis
- create categorical variables using cluster analysis

The software tackles these challenges by first containing all the mathematical techniques as part of the overall software. The analyst needs to interpret these results and assess what the impact is on the business. In addition to the ability to conduct high-end mathematical routines, capabilities should also exist that allow the analyst to create new derived variables. The ability to create derived variables is an important skillset within the data science discipline. At the same time, software is only a tool as the identification and creation of derived variables relies on the insight and knowledge of the data science practitioner.

From a more general or philosophical perspective, the guiding mission behind data science software is about automation. This is not concerned with automating the data science discipline itself, but rather those tasks that yield minimal critical thought. Instead, software advances now allow the analyst to devote more critical thought to the problem and to consider different approaches and techniques that might help solve the problem. The mundane task of creating programs to generate the necessary information for critical thinking to occur is now automated. An example of this is the development

and evaluation of models. Once the analytical file is created, multiple models can be very quickly created without requiring any programmer resources. Using a GUI interface, the analyst, through selection of the appropriate technique, can literally develop a model on the fly. Gains charts or decile tables are automatically produced to evaluate results of the new model. This factory-like development of models allows the analysts to look at many different options for a given model. More time is now spent by the analyst looking at different modeling options rather than programming results. The analyst still needs to understand in detail the component or variables of the model as it is incumbent on us, as data scientists, to understand the details of any "black box" solution. Validation of model results in any business initiative needs to also occur. But how does this occur effectively in an automated modeling environment where hundreds of variables are created? Automated routines now look at the KS (Kolmogorov-Smirnoff) statistic or the volume of that area of the curve between the flat line and the parabola (lift curve), which has been discussed previously (see Fig. 13.4).

If the change between the KS statistic (the area under the curve between the parabola and the straight line) in model development and its current application fails to reach a certain threshold, as determined by model development, then the routine will trigger the user that the model has significantly eroded. Organizations, such as SAS, offer specific tools in this area (such as Model Manager) to help identify when these triggers occur.

Evaluating Software

The first and primary considerations are the actual software solution's results. Are better results produced with other software options, or with the existing software? As we saw above, looking at the KS statistics between different software solutions represents one option in evaluating different software. Another option is to simply compare how a given solution rank-orders the results, which is just simply another way of viewing performance. If the primary data science objective is about the differentiation of results, then comparison of results across different software solutions might yield the following graph (Fig. 21.1):

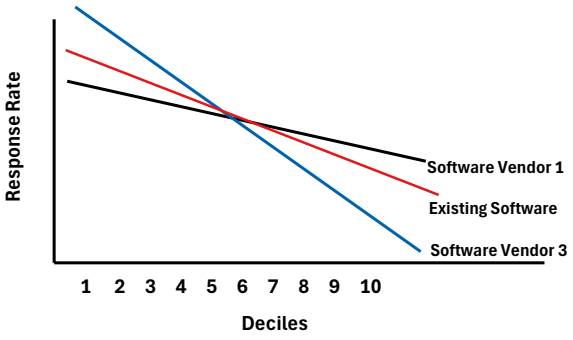


Fig. 21.1 Evaluating different software solutions using Lorenz curves

In our above example, Software Vendor 3 would be the optimal choice as it provides the best rank-ordering results from the top decile (Decile 1) to the lowest decile (Decile 10).

Another factor in evaluating software relates to the ability in creating derived variables. High-end users are typically programmers who can manipulate source data from a number of files and then create new variables. The ability to create derived variables with minimal programming saves programming time, but also empowers other users who may not have a programming background. In this case, the user would use the appropriate interface and modules as pointers by being able to do the following:

- Identify source files and fields that will be used in the exercise
- Link appropriate files and use appropriate match keys in linking these files
- Mathematical and statistical routines to summarize and manipulate data into other variables

Even though software may facilitate these above functions, most data science practitioners will rely on their hard programming and technical skills for these above tasks.

Besides being able to derive new variables, the software should also have flexibility in terms of looking at a variety of statistical solutions. Both linear and nonlinear solutions should be examined with such techniques as:

- RFM
- Linear regression
- Logistic regression
- Neural nets
- Genetic algorithms
- CHAID/CART

Scalability as a Consideration

Scalability is an important consideration since expanding organizations will presumably have expanding data requirements. This is, of course, important within the world of Big Data. Questions should be asked as to whether the software has specific limitations regarding the amount of data that can be used in any application. Additionally, the user should also be cognizant of the fact that certain mathematical routines are much more sensitive to increased volumes of data. This is indeed the case with the application of various AI-type mathematical routines.

Portability of Output as a Consideration

Often enough, the output of certain applications needs to be input into other applications such as Excel, Word, or PowerPoint. It is important to consider how easy is it to output data to more user-friendly business applications. In the early days of data science, this task was quite onerous as it often required much cutting and pasting, as well as the typing of raw output from a statistical application into a more user-friendly application such as Excel. In today's environment, raw output can now be sent directly to Excel. This has resulted in significant time savings since data scientists now spend more time on analysis rather than massaging the raw output for presentation purposes. Besides the portability of output to other applications, the inputs into the application need to be considered. For example, how compatible is the software in receiving data from a variety of different formats such as comma delimited, tab delimited, SPSS datasets, SAS datasets, as well as object-oriented or semi-structured data?

Reports and Diagnostics as a Consideration

The ability to produce the right reports and statistical diagnostics are obvious deliverables in producing the optimal solution. This is best achieved by software applications that have flexible reporting capabilities which are easily generated by the end-user. Alongside these reports, functionality will also exist that provides the creation of standard reports with minimal user intervention.

The use of statistics in any applications should provide sufficient diagnostics that allow the user to mathematically compare different solutions. For example, the use of R^2 in linear regression or concordant/discordant pair ratios in logistic regression represents specific metrics that would be used to compare different statistical solutions from a mathematical perspective.

Yet the most important reports in these software applications are those that can yield information about the business impact of a given solution. In building models or targeted solutions, this typically means the generation of gains tables or decile tables, which have been discussed at length throughout this book.

Gains charts, as seen previously, demonstrate how well the response model applied to a given campaign differentiates segments (deciles) of customers based on response rate. At the same time, revenue numbers and cost numbers can be used to convert the response rate performance into an ROI metric. The ROI numbers can then be used to select customers that would be most profitable within a given business initiative.

Looking at Software from the Business Analyst Perspective

As we observed in the small section devoted to dimensions and metrics in this book, analysts now have the capability to conduct their own analysis. These analyses do not require the intervention of a high-end user such as a statistician or data scientist. Instead, the analyst is simply trying to obtain top line results to a previous campaign or business initiative. Generating results to various campaigns, or conducting various ad hoc analyses, does not require sophisticated analysis. Instead, the user is required to create crosstab type reports that can yield the necessary information and insight on a particular business problem. The user does not need to write programming code to generate reports; they simply use the GUI to generate the appropriate reports. The user, however, does need to understand certain mechanics of the data, such as what information is to be measured in terms of performance compared with how they want to view these performance measures. Examples of this might be a report on income (performance measure), but providing views or dimensions of this report based on gender and region.

This kind of empowerment does require the purchase of certain software, often referred to as business intelligence (BI) software. Without getting into too much detail, which can be found in many database technology textbooks, the technology behind this type of software deals with the concept of creating data cubes. These cubes, in theory, represent pre-built queries of the data where the design of these cubes attempts to address all possible queries of the data.

In assessing this type of software, the primary concern should be with ease of use. In other words, how quickly can an analyst get up to speed in using the software? Ease of use can translate into more users having access to data. This enhanced empowerment to end-users means that more analysts are dealing directly with the data, rather than having to go through a programmer to generate reports. Ultimately, this will lead to better decisions due to quicker turnarounds of results and information.

In addition to ease of use, the other key consideration is the range of variables that one can use in designing reports. Obviously, a more extensive range of variables will provide enhanced flexibility in being able to generate reports.

In choosing specific software, considerations will vary in importance according to the organization's needs. For instance, an organization just commencing activities in analytics may look at simple tools, such as BI software, which tackle only a limited number of variables but, most importantly, emphasize ease of use. A more advanced organization around analytics may place reduced emphasis on ease of use and more emphasis on the variety of statistical tools which are available within the software. As with any purchase decision, cost will always be one of the determining factors. Its weight in the overall purchase decision, however, will be contingent on the importance or weight of other considerations within the decision-making process.

Choosing software for an organization is no easy process. It requires a deep understanding of the analytical needs of the organization. By understanding these needs, the appropriate weights can be allocated to those key considerations in making the right software selection. As is the case with most purchase decisions, one size does not fit all. Instead, the software choice for the organization needs to address its customized analytical needs.

Key Takeaways

- The discipline of data science needs to be integrated into the structure of an organization.
- Typically, there have been two approaches:
  - A more centralized approach with the chief analytics officer (CAO) as the lead.
  - A more decentralized approach. where data scientists report directly to the functional head of a given department.
- A third approach is the hybrid which encompasses both options.
- People are the most valuable resource in data science, and there are a variety of ways for the data scientist to become better educated.
- Software tools are now more accessible and facilitate the many data science activities.
- AI and GENAI is a game changer. It will transform the role of the data scientist, as will be discussed in later chapters.

CHAPTER 22

Social Media Analytics

By putting a different spin on Shakespeare's famous quote from Hamlet, "To Be or Not to Be", the question of analysis, and what it means within the social media sphere, is an extremely relevant issue. Because of the very nature of the medium, privacy implications enter a whole new realm of consumer sensitivities and perceptions. What do I mean by this? The more traditional concerns of privacy have focused on the fact that a consumer has had some engagement or interaction with an organization. The information concerning this engagement or interaction is captured and can be potentially used for marketing purposes. The laws around PIPEDA (Personal Information Protection concerning Electronic Documentation Access) are quite clear with regard to how marketers can use information regarding the transfer of information between the consumer and a given organization. But the world of social media raises the debate to a whole new level. Consumers are now interacting and engaging with a variety of individuals and potentially organizations within their community. For example, I may be a valued CIBC customer who has responded to an ad on Facebook regarding some RRSP promotion. The information pertaining to my response would certainly be captured by Facebook, which would also have access to all the other data pertaining to my interactions within my social network. For instance, my plans for Christmas, which would certainly be discussed within my forum of friends, might be of extreme value to a CIBC marketer in promoting certain products to me at that time of year. Remember, I am a customer of CIBC, and they are merely attempting to service me in the best manner possible. Yet, the information that they may be using in determining the right marketing approach to me, might in fact be generated from information pertaining to my Facebook interactions. Currently, the challenge in using Facebook data would be in linking this individual-level data back to the same individual but as a bank customer. Even if this linkage could be accomplished, however, it is highly unlikely that a bank or other leading ethically oriented

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_22

business organizations would contravene the trust of its customers by using information that the customer expects to be within the confines of his or her social network.

The number of social media platforms continues to expand, providing a variety of new outlets for information about the individual. This is nirvana for the data scientist, who can simply develop better solutions with more data and advanced technologies such as artificial intelligence (AI).

To address this issue, many organizations now have privacy policies that have a specific component dealing with digital and social media data. From a public policy perspective, governments today are updating their privacy policies to reflect not only an increasingly digitized world but also one that is more focused on the use of AI.

Presuming, of course, that we can address these privacy concerns, the ability of organizations to understand the social media behaviors of its customers simply means more effective marketing. As with all data, the more significant benefits are conferred when we can analyze all behavior at an individual level as well as actioning the learning at an individual level.

From a targeting standpoint, additional variables can be created; these can be included as potential variables in any predictive model. Besides the targeting benefit, there are unique insights on how to communicate to these customers which is provided by this medium. For example, one key finding in any social network is to determine or identify the influencers versus the followers by analyzing the network activity of each record. Data such as the number of contacts or friends within their network, and the number of times that they reach out directly to their friends, as well as the number of times that friends reach out to them along with other types of activities can all be used to identify influencers vs. followers. By using network analysis, each customer has a network of activities that have been made within a given period of time. The key in determining whether or not an individual is an influencer or a follower is through examination of both the breadth and depth of that individual's network. Here, breadth refers to the number of contacts within the person's network, while depth refers to the level of engagement that each person has with all their contacts. Various metrics and indices can be calculated which allow the analyst to create an "influencer score". A component of this type of analytics exercise would be to determine the threshold or benchmark from these metrics/indices that classify someone as an influencer vs. a follower. This, of course, will vary from exercise to exercise.

Once we have determined how to stratify customers into influencers vs. followers based on their social media behaviors, marketing communication strategies could be formulated that lead to different types of messaging to both groups. Relying on the old RFM (Recency Frequency Monetary)-type approach which is used in the more traditional CRM (Customer Relationship Management)-type channels, marketers can create key similar type behaviors within social media. For example, habits related to average length of time for a given session, the most common time of day when they logged on and how

often they log on within within a given period of time can all be used to derive insights regarding consumer purchase behavior. This pseudo-RFM type behavior can be examined for both influencers and followers. Furthermore, this behavior can be also further analyzed by exploring the different user groups to which the individual belongs.

Another challenge in social media analytics is the issue of marketing attribution which attempts to determine how much return on investment (ROI) from a given marketing campaign can be attributed to social media. There may be no right answer or panacea to this challenge. Unless there is some direct method whereby a given person logs onto a certain web page and registers for a given product or service within the social media campaign, it is very difficult to determine direct dollars that can be attributed to social media. Yet, even though this functionality exists, one could argue that the non-registrants participating in a Facebook contest might be contributing to ROI by purchasing at a store, through TV, or via direct mail. These same kinds of challenges exist for the mass advertising industry, except that the key advantage with social media is that the denominator in determining the number of people that clicked onto a fan page or contest is known. Given these limitations, all the analysts can do is to provide general direction in whether the campaign created additional engagement, and whether or not this additional engagement translated into additional dollars. On the engagement side, metrics such as the number of people logging onto a fan page or contest and how long they are on the site can represent some kind of proxy and comparisons of these metrics can be made to other social media campaigns to determine the level of success. By the same token, the anecdotal analysis of social media campaigns over time can determine the overall impact on incremental sales. The classification of campaign periods into high, medium, and low social media marketing allows executives to see the general impact of social media marketing on sales. However, one cannot directly attribute the ROI back to a specific social media campaign. All that can be concluded is that the campaign generated an X% improvement in engagement relative to other campaigns. The leap of faith for executives is that X% improvement in engagement translates to incremental dollars, but the precise number is unknown.

Assuming that privacy concerns are addressed, the more granular-type analysis, as evidenced by being able to capture social media behavior at the individual level, is always going to be the preferred course of option for marketers. But does that mean meaningful analysis cannot be done unless we are able to overcome these privacy restrictions? Of course not.

Analysis can be done on all the records within a social media network without knowing the unique identity of the specific record. Once again, I can identify influencer vs. follower records, as outlined above. The difference here is that I do not know the identity of the individual record. The social online habits can all be examined for both influencers vs. followers as well as the different user groups that exist within the medium. The difference here is that targeted marketing to individual online users cannot occur. Yet, specific

communication strategies can be developed as insights can be drawn around different user groups as well as influencers vs. followers. Rather than deliver one message for all users, specific targeted messages can be created within a specific social media site and, potentially, to different user groups. But the message itself is not personalized to the distinct person's identity.

As social media within the marketing area continues to evolve, the analytical possibilities will expand, recognizing, of course, that privacy concerns must be addressed. In addition to a better understanding of how social marketing fits in within a given organization's marketing strategy, marketers will also need to have a better grasp of analytics and how that fits into a given social marketing strategy.

Key Takeaways

- Exponential data growth arising as a result of all the social media platforms that are now readily available to consumers.
- The challenge, as seen in even the non-digital age, is how to link all this data back to the same individual.
- Social media analytics can be done to better understand followers vs. influencers through network analysis.
- The online habits of social media users can be utilized to craft messages to user groups that are not personalized to the person's identity.

CHAPTER 23

Credit Cards and Risk

Fair Isaacs

The birth of credit cards in the late 1940s and 1950s necessitated the demand for techniques and processes that reduced overall credit losses. In some sense, the credit card industry was the industry pioneer of data science.

The bulk of this responsibility resided in the credit operation and collections area of these organizations. Business rules were the norm, such that if a person's credit criteria reached a certain level, various procedures were introduced to deal with these people in order to reduce the likelihood of them going into default. Examples of these procedures could range from a phone call indicating the organization's concern with a client's current credit risk status to shutting down the account completely until payment is received. These procedures were successful in terms of reducing overall credit losses. However, given the actual amount of these losses on their books (which, in many cases, was estimated to be in the tens of millions of dollars), senior company executives were always considering other techniques and approaches which could reduce these losses, even marginally. A 1% improvement on a book of $50 billion losses translates to $500 million in annual direct savings to the bottom line. With the ability to reduce losses at this scale in perspective, companies began to consider other options or ideas which would reduce their overall risk exposure.

One technique which was considered as a business tool was the use of statistical techniques (namely multiple regressions) as a way of targeting high-risk customers. During World War II, the use of statistics was first used to predict certain enemy behaviors and tactics based on prior history. The success of these measures led company executives to consider it a potential business tool.

The need for these types of tools led to the formation of a company called Fair Isaacs. Using mathematical techniques such as multiple regression, the company created a scoring system which created risk scores based on the individual's prior credit behavior. In this system, individuals who applied for a

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_23

credit card were told that their credit history would be used to determine whether they would be approved for a credit card. At the same time, these individuals were told that should they be approved, their ongoing credit information would be periodically assessed.

For new applicants, the process of determining approval involves comparing an individual's score against a selected cutoff range. Individuals who scored above the cutoff were accepted for the card, while those who scored below were declined.

On an ongoing basis, credit card customer-related information was sent to Fair Isaacs, where updated scores based on the individual's current credit history were attached to the customer's credit card record. These scores were not meant to replace the current credit business rules and procedures that handled high-risk customers; rather, their intention was to augment them. With the scores, the credit card companies had additional information which ultimately increased their decision-making capabilities. These developed skills led to a refinement of the current credit business rules and procedures concerning the treatment of high-risk customers. As with any sound organization, the introduction of these new rules and procedures was compared to the status quo to determine the overall net impact to the business. Following the evaluation, the new rules were implemented and proved successful as Fair Isaacs is still in business today; indeed, it is still regarded as the pre-eminent experts when it comes to credit-risk modeling. In fact, the power of their capabilities is best exemplified by the longevity of their models. In many cases, the models and their scores have been in use for five years without any major updates. In some cases, the longevity has been as long as ten years.

The success of Fair Isaacs with the use of these techniques led other organizations to bring such operations in-house. Recognizing the huge value of loss prevention, even at the margins, organizations hired their own statisticians and mathematicians so that some mathematical expertise could be solely devoted to their business. Organizations in today's workplace will produce their own tools, in order to further enhance the decision-making capabilities of their credit operations.

In Canada, comparable organizations, such as Equifax and Trans Union, offer credit bureau services as well as mathematical services using scores. The Equifax "Beacon Score" is a commonly accepted credit-risk score used by companies to assess a customer's creditworthiness. Besides the actual scoring of customer records, these companies have conducted activities around the collection of credit data regarding customers and have created huge customer databases. This lucrative activity allows these organizations to conduct what is called a "credit bureau" on individual customers. Given that the customer has given their consent to do this, any financial institution which is offering a lending-type product can obtain a "credit bureau" on a potential applicant. This "credit bureau" is essentially a copy of their record which details all their current credit history. Using key information and fields from the customer's record, it is quite common now for companies to use a combination of credit

score and some credit history information to make their credit approval decision. One example of this might be that credit card applicants must have a credit score above 650, and have gone 60 days without making a payment no more than once in the last 12 months.

In fact, the success of these credit scores in assessing the likelihood of credit default has crossed into the area of business to business (B2B) marketing. Companies such as Dun and Bradstreet have offered this service for many years. Paydex scores (a measure of the company's likelihood of going into default payment) is used as a key metric by most organizations to determine whether to extend credit to other organizations. In addition to the Paydex score, Dun and Bradstreet also collect a vast array of information on payment history, which is transformed into indexes by industry sector. This means that the index score for the number of times that a company has paid its bill in over 60 days in the last year for Company "A", presuming that it is in retail, can be used to determine how far above or below the company is in relation to the sector average. Besides credit payment history, which is collected and reported at an industry level, there is also information pertaining to the demographics of the company. Information relating to company size, tenure, sales, industry sector, etc. is readily available for organizations to use in their business decision-making processes.

Credit Risk and Marketing Behavior

Although the pioneering techniques of data mining began within the credit-risk industry, the extension of these techniques to the marketing realm of the business world has also resulted in marketers using credit-risk information as key components within their own decision-making processes. The use of such has been shown to be a powerful predictor of consumer behavior. From a marketing standpoint, this makes sense in that the underlying credit patterns of a particular individual will influence an individual's decision to purchase a particular product or service. In building models to predict response, or the likelihood of someone filling out a credit card application, it was discovered that the most responsive prospects were those individuals who were also more likely to seek more credit. These credit seekers were indeed high credit-risk-type customers. This information was confirmed when these initial response models performed very well in getting prospects to fill out credit card applications. However, average approval rates (which involved looking at the credit bureau behavior of these prospects and making decisions to either decline or approve them) decreased dramatically with the introduction of these models. The next step was to build models that optimized both the likelihood of filling out an application and the likelihood of being approved.

Beyond the credit card example, there have been other cases of strong relationships between marketing behavior and credit risk. Within the auto and property insurance industry, the likelihood of purchasing policies is positively impacted by good credit behavior. From a claims-risk perspective, higher

credit-risk customers are more likely to have a claim. Meanwhile, the retention behavior of insurance customers generally seems to indicate that a higher credit-risk customer is more likely to defect.

Trigger-Based Behavior and Fraud Risk

In the world of data mining, one of the new terms that is more commonly spoken of as a key concept for marketers is the notion of trigger-based marketing. Trigger-based marketing represents the identification of key marketing behavior which is out of pattern. For instance, a customer typically spends $100 per month; then, within the last three months, they have been spending $500 per month. Clearly, there has been a change in customer behavior. This type of information is invaluable to marketers, as the ability to develop products and services around this change will always be a marketing priority. There have been a number of examples of this trigger-based marketing that have been used by companies. One of the most common of these is life-stage marketing.

This approach works as follows: Specific services and products are required depending on the various stages of life and the situation of the consumer. Since most companies have hundreds of thousands of customers, the challenge is how to individually identify the life stage and situation of each customer. The use of a marketing database can certainly facilitate this process, but it is likely to be less than perfect. For example, key information, such as age, gender, number of household occupants, education, income, occupation, etc., may or may not be contained in a database. Even if these fields are present, it is highly likely that most of these fields will contain very large proportions of missing values. Determining the precise life stage of a particular consumer will depend on the data. Given the data environment, broader assumptions regarding the life stage of a consumer would be inferred by the marketer. Let us consider a few examples. A marketer may look at a group of people over 60 years of age. They might look at their spending patterns in relation to a company's services and products and determine that there is a group that have decreased their spending by over 50% in the last two years, but who are maintaining this lower level of spending. The significance of this is that there has been a major decline in spending, but that this decline represented just a one-off event rather than continuous declining behavior which would, ultimately, lead to outright defection. By identifying this group of people aged 60+ who decline but maintain their spending at this new lower level, the marketer may infer that these people represent retirees. Although this might not be the exact situation for the entire membership of this group, marketers could design a communication strategy which recommends specific retirement products and services, but without directly calling the consumer a retiree.

Another example might be a customer who is married and who ultimately increases their spending. Further analysis of their spending may indicate that much of this new spending is attributed to products related to children. Once

again, this type of knowledge can be used to develop specific strategies around this lifestyle change.

Although trigger-based marketing is a relatively new phenomenon, the principles behind its identification are well established. For example, fraud models have been around since the inception of credit cards. Businesses have recognized the tremendous benefit of detecting fraud and its ultimate prevention. The ability to do this was based on being able to detect irregular spending habits and thereby to create a flag on the account at point of sale (POS). This means that when the card is used, a flag notifies the merchant at POS that further investigation is warranted before accepting the sale. Further investigation might include a call to the credit card company to review the situation. An important aspect of creating this fraud flag, however, is the ability to detect the trigger or "out of pattern" spending. This trigger will vary from company to company, yet, in principle, the goal is to identify actions that differ significantly from the norm. The use of simple variance analysis might be one approach where behaviors are examined at different numbers of standard deviations from the norm. Validation of this approach would occur in assessing how many frauds are identified with the current set of rules. By employing some sensitivity analysis, the % of all actual frauds captured can be identified using different sets of triggers (number of standard deviations). Meanwhile the fraud rate looks at the number of actual frauds divided by the number of observations in the sample that are outside the number of standard deviations threshold (Fig. 23.1).

From this example, a good cutoff rule for identifying an appropriate trigger might be one standard deviation. Below one standard deviation, marginal improvement occurs (increase of only 10% of all frauds), yet a decrease in fraud rate of 25% is observed.

Besides using variance analysis, which might be the quick and easy approach, one could use the more robust predictive modeling approach. This approach identifies records that were fraud vs. non-fraud within a particular period, and then identifies the particular demographics and behaviors of those records prior to the post-period (pre-period). Constructing the information in this

# of standard deviations from norm	% of all actual frauds captured in sample	Fraud Rate
0.5	90%	45%
1	80%	70%
2	50%	75%
3	25%	100%

Fig. 23.1 Example of fraud capture at different standard deviations from the norm

manner enables the individual to identify those key behaviors and demographics which will predict the likelihood of a fraud event occurring in some defined post-period. The determination of these periods (pre- and post-) is critical to the success of these types of models. For instance, the post-period probably represents the time at which the fraud was actually reported by the customer to the company. The pre-period in this case would be very short, perhaps consisting of the last month of spend prior to the fraud being reported. In the case of non-fraud events, the pre-period would be the current month of activity. From an intuitive standpoint, it is known that the critical fraud behavior occurs just hours prior to the detection. One could ask why the benchmark isn't several hours prior to the event. The problem here is that credit card spending, by its nature, can be very volatile due the huge variety of merchants and establishments that accept the card. There needs to be an established time horizon which can smooth out this volatility. Therefore, this behavior needs to be benchmarked against a time window which is a reasonable representation of that person's purchase behavior. In other words: is it rational to expect customers to use their card at least once per month? Or do they use their card either more or less frequently than this? Typically, good customers can be expected to use their card every month and accordingly, based on that learning, one month can be used as the spending benchmark for fraud detection. The actual threshold for determining whether or not a record is fraudulent would be based on a score that yields a rate of fraud which is unacceptable to the organization. In Fig. 23.2, this threshold is highlighted in bold as being the cutoff point where all records above this fraud score are considered fraudulent.

Once again, like any approach, the solution, and any resultant decision-making, can be validated based on how it performs.

Fig. 23.2 Decile chart of fraud model

% of Customers Ranked by Fraud Model Score	Minimum Fraud Score in Interval	Average Fraud Rate
0-10%	0.051	4.40%
10-20%	0.045	4%
20-30%	0.042	2.90%
30-40%	0.038	2.70%
...		
90-100%	0.0045	0.50%

Key Takeaways

- The credit card risk industry, and in particular Fair Isaacs (FICO), were the real pioneers in the utilization of data science for more effective business decision-making.
- The use of credit card-risk data has often been used by marketers in their targeting tools.
- In assessing risk, simple techniques, such as out of pattern spend, have been used as one tool in targeting high-risk customers.

CHAPTER 24

Data Science in Retail

History of the Retailer

When considering consumers and their behavior at its most basic level, typically the transaction event which occurs between seller and buyer is the first thought. The notion of being able to sell a good which is desired by a consumer represents one of the basic tenets of consumer behavior. In earlier times, this was best observed during open forums where buyers and sellers converged at a public venue. The transaction exchange during those earlier times involved the bartering of goods whereby both buyers and sellers exchanged roles during the transaction. As the transaction evolved over the course of time, goods and services could then be exchanged with some form of common currency. With the advent of currency as the primary mechanism for the purchase of goods and services, more formal structures such as the store became the primary institution for this type of exchange. The typical "Mom & Pop" store represented the early type of stores, which attempted to serve a local community with a limited population. Historically speaking, this was the norm as much of society was situated in primarily rural areas. Remnants of the "Mom and Pop"-type store are visible today in the form of convenience stores such as Becker's and 7/11. In these stores, the consumer is often personally known to the seller. The song lyric, "Where everybody knows your name", the theme tune of the TV program *Cheers*, which was set in a local bar in Boston, encapsulates this idea. Within these "Mom and Pop" stores, CRM was a particularly concrete notion. This made one-to-one marketing very straightforward.

With the explosive population shift towards urban and suburban areas, however, these "Mom and Pop"-type stores have evolved into larger retail organizations, which are familiar to this generation. In this type of environment, the typical one-to-one marketing/selling approach of the earlier types of retail outlet is no longer observed. In its place, mass marketing processes are promoting products and services targeted at an ever-expanding population. However, this

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_24

mass marketing approach, despite its continued popularity within retailing, does have considerable limitations. Retailers are facing ever-increasing competitive pressures, in a wide range of areas: price, access to information by consumers, loyalty programs, e-commerce capability, and, perhaps more importantly, less patience on the part of the consumer to hear the retailer's message, given all the other demands on people's time.

Retailing Today

How have retailers adapted to this new paradigm? Recognizing that data and analysis is the key to better understanding of, and ultimately providing better service to, the customer, retailers have investigated processes and procedures which strive to achieve this objective. The challenges for retailers are somewhat unique, especially in comparison to banks or insurance companies. Banking and insurance transactions are gathered electronically and at the individual level. The data or information is typically captured at an account ID level, which can then be summarized at a customer level. For example, a client's deposit, withdrawal, borrowing, investment, and credit card behavior can all be viewed in a holistic manner in order to develop meaningful marketing intelligence about their overall banking practices. Similarly, marketers within insurance companies can analyze the demographic information alongside the prior claim and premium information of a given policyholder to derive meaningful marketing intelligence concerning a particular group of policyholders. The ability to do this, however, is contingent on the fact that all behavior is captured in a policy number; this is similar to the banking example where information is captured through an account ID. For many retailers, there is no capture of information at the customer level.

Retailing Challenges

For the retailer, the capture of information at the individual is often not that simple even with loyalty programs and e-commerce capabilities.

Suppose retailer XYZ has a loyalty program containing 200,000 customers. Customer transactions regarding the loyalty program are typically captured using a card which is presented by the consumer at point of purchase. Suppose a member goes into a retailer and uses the loyalty card to purchase a $100 pair of tennis shoes because the salesperson prompted them to use the card by asking if they had one. Now suppose that, one week later, the same member goes into a different store and decides to use their AMEX card and spend $200 on a jacket. On this occasion, they forget to use the card because they were not reminded by a less loyalty-conscious salesperson. The AMEX transaction, card number, amount, and date, would be captured because this information must be transmitted back to American Express for billing purposes. How can retailer XYZ use this Amex information at an individual level? The answer is that it can't. The only information that can be used at an individual level are those transactions which can be attributed to the loyalty card.

E-commerce transactions face the same difficulty in that the capture of data is confined to the e-commerce platform. Again, we have many customers who will purchase at point of sale (POS), meaning that their information cannot be tracked back to the individual level.

How Can Non-individual-Type Data Be Used for Retailing?

But what about these POS transactions that do not use any loyalty-type card. That is, those in which no information is collected on a database. Is this information still useful? The answer is yes, when viewed from a different perspective. In the Amex case, the members' behavior can be grouped with other Amex transactions at the store level. Store statistics, such as the % of all transactions due to Amex or the % of total amount due to Amex can easily be created and then produced based on certain time periods, all at the store level. This same approach can also be used for other types of payment vehicles such as MasterCard, Visa, cash, and now e-commerce. In the case of cash, there may be no electronic recording of the transaction as some cash transactions still occur through the exchange of currency or via debit card. Despite this information not being recorded, it is still possible to approximate the level of cash transactions through default since the total transactions and total credit card transactions, as well as the e-commerce transactions, is known within each store. In this case, the total cash transactions for a given store are simply the difference between the total transactions and the total non-cash transactions.

With this information available at the store level, it is possible to identify those stores that are high value. In order to have a true calculation of this number, one would not want store size to be the main determinant of value. Obviously, a metric that calculates the average sale value per square foot would be a better representation of true store value. Having identified stores that are of high value, it is then possible to determine whether or not payment type is a key determinant of a store being high value.

Using Stats Can data, trading areas around these high-value stores can be defined, and Stats Can demographic characteristics that pertain to these areas can be overlaid. Analytics is then used to identify those key demographic characteristics that differentiate a high-value store from a regular store. This information could be used in several ways:

- Targeting of prospects to become customers that live in these high-value trading areas, or areas that look like high-value trading areas.
- The opening of new stores in trading areas that look like high-value store trading areas.
- Potential store closings based on the fact that they are located in low-value trading areas.
- In-store programs that are developed using the key demographic characteristics of customers that live within the trading area of a high-value store.

Despite the information being available only at an aggregate level, marketers can use this vital information to help develop store programs and strategies as well as acquisition programs to attract new customers.

How Can Loyalty Card Information Be Used?

As previously stated, information accumulated within any loyalty program can be used at the individual level, with more robust techniques such as predictive modeling being deployed for this group of customers. Cross-sell/upsell programs, along with retention programs, can all take advantage of customized predictive models which can be used to identify key customer segments for a given customer loyalty program.

When conducting these types of analytics, an important consideration is whether this loyalty card behavior represents their total purchase behaviour. In other words, are the insights from loyalty card customers significantly different from those related to general customer behavior. The use of market research would be one possible method to attempt to better understand the impact of this gap. Yet, data scientists are very quick to observe that what consumers state in a market research survey is often very different from what they will do with the "do" being captured in a database. Given this dilemma, the challenge is to apply these data science techniques to loyalty card customer behavior, always recognizing this limitation. Rather than spending too much time trying to develop strategies around this gap, marketers have come to the realization that they are better off in dealing with what is known. In this case, the loyalty behavior of a given customer should be the focus of the marketer. Thus, even though a client may, in fact, be the best customer of retailer XYZ, they are a low-value loyalty card customer and should therefore be marketed as a low-value customer within the context of any loyalty marketing program. In this respect, their overall high-value behaviour is simply irrelevant as it cannot be effectively used within any CRM-type program.

Using E-Commerce Information

Although this rich digital purchase information is extremely powerful in terms of understanding individual e-commerce behavior, the same loyalty limitations expressed above still apply in this case. There are still a significant portion of non-loyalty and non-ecommerce customers that comprise significant retail purchase information. This essentially implies again that analytics should be customized separately towards loyalty customers and e-commerce customers.

Using Information to Develop Product Strategies

Retailing as a discipline tends to focus on product movement. The ability to quickly turn over inventory and establish the appropriate inventory control procedures are the primary objectives of any retailer. The use of data science can be a force in achieving these objectives since retailers strive to attain insights that can be used to develop product strategies within a given store. But what kind of information is available to the data scientist, and how is it used? Information at an individual level represents the optimum scenario for data science and analytics since the use of statistics is better leveraged through this increased granularity of information. At the store level, information can be gathered at an individual level, but the store record itself is at the transaction level rather than the customer level. In this exercise, the actual transaction itself becomes the focus of the analysis as retailers strive to better understand what occurs within the transaction event itself. Retailers will want to know what type of products are bought within a given transaction event. More importantly, are there certain types of products that appear to be bought together? And do these product relationships differ by type of store?

Observation by staff members, complemented with store report statistics, yields valuable information on what products are being sold and, more importantly, changes in products being sold within a given time period. However, what will be the case when the store is selling hundreds of products, and hundreds of thousands of transactions are occurring at a monthly level? Can this information still be considered reliable? The answer is yes; however, data science can help to supplement this knowledge by using more science to uncover patterns that might not be observed either through anecdotal observations or through standard store reports. Through rigorous analysis, data science can help to identify unique product patterns that occur within a given transaction. For instance, say product A is purchased; what might be the next most likely product purchased within a transaction event? Using product affinity and basket-type analysis, retailers can develop product bundling strategies within a given store. These strategies might be simple, such as placing products with high affinity near each other alongside in-store promotions that collectively market a particular bundle of products. At the POS, it might simply be the cashier reminding the consumer about another product based on the existing purchase of another product.

An important aspect of success in data science is treating information as a valuable corporate asset. This is certainly no less important for the retail sector. Yet this success can only be realized if retailers recognize that information is not only an asset but also a critical competitive advantage. With this kind of mindset, retailers can truly differentiate their products and services within the competitive marketplace, which is ultimately what all retailers strive to do.

Campaign Management System

In this case of a prior client that we worked with, the client achieved relative proficiency in being able to use predictive modeling techniques to better market their insurance products. The company, a retail organization, marketed a variety of insurance products sold by various suppliers but marketed against the retailer's customer database. This rich customer database became a very competitive property as each supplier wanted to market their products to the best names as determined by the model. If no rules were in place, each supplier would be ultimately promoting the same names to obtain the best results. Within a very short time, the ultimate consequences would be list fatigue, as well as increasing customer complaints followed by attrition. To alleviate this situation, the company instituted a set of arbitrary rules whereby each supplier would be allowed to promote a given name only once within a 12-month time frame. However, was this restriction reasonable? Without any analysis of the case, this question would remain unanswered. The retailer understood that, to answer this question, a campaign management database would have to be developed. In other words, the organization needed to record information at the customer level pertaining to the frequency of promotion, the timing of promotion, and the type of promotion. Alongside promotion, they needed to keep track of the outcome of the promotion effort (i.e. did the customer respond, or what was the non-response outcome, if any?). The challenge, then, was to build this campaign management database which would help to determine the right number of promotions, the timing of promotions, the type of promotion, and the actual promotion/campaign outcome which would optimize campaign performance.

The priority, then, was to conduct a preliminary needs analysis whereby detailed knowledge on the current data environment could be obtained. Having built models in the past, a strong base of knowledge regarding the data environment is essential to the process. However, it was still important to understand the data process in terms of how data was transferred between the retailer and its insurance partners. This meant that when a campaign occurred, the following needed to be understood:

1. What were the steps or activities that were required in creating lists after scoring the names with the various insurance models?
2. What data-processing activities occurred with campaign feedback?
    - How are respondents tracked, and how are they fed back into the retailer's database to determine purchase activity?
    - Is non-responder activity being captured? If so, how is it being recorded? This is particularly relevant for telemarketing programs where the telemarketer will record the non-responder outcomes such as wrong number, call again, do not call again, etc.
3. Is the above campaign activity being captured historically?

In points 1 and 2, the detailed investigation enabled the production of a project plan and data flow chart. More importantly, it also outlined how data would flow into and out of the campaign management database. However, in point 3, it was observed that no prior campaign activity was being captured. Essentially, this database had to be created through the creation of a campaign management database. This required the identification of the appropriate information which needed to be captured within the database. Second, how the information should be organized within the database needed to be outlined. The major information components which were required to be captured in this database were as follows:

- Model behavior
  - Promotion behavior
  - Disposition behavior

With this above information, it was now possible to determine how customer behavior was impacted by the following:

- Changing model behavior
- Recency, frequency, and type of promotion
- Changing customer disposition activity

Because speed of processing was not an issue, the organization structure was very basic in that simple database tables were created to store this information. Yet, although the design was simple, the value to the client was designing a process that allowed the retailer to achieve the business objective of being able to make better decisions based on a customer's prior campaign history. The results within nine months of this launch revealed the following (Fig. 24.1).

Figure 24.1 indicates how the institution of this system improved overall results. Using the key metric of average premium per customer, a metric can be created which determines how many months it would take for a marketing program to break even. Obviously, a program which generates few responders and a low premium would require a longer time before it paid for itself. In this chart, the time to pay back is approximately 50 months prior to any modeling

Campaign	# of Leads	# of Sales	Total Cost	Cost/Sale	Avg. Prem/ Cust./Month	# of Months to B/E	
1	20,000	285	$32,000	$112	$2.10	53	
2	20,000	303	$32,000	$106	$2.34	45	Pre Modeling
3	40,000	1,134	$64,000	$56	$4.17	14	
4	30,000	1,029	$50,000	$49	$4.40	11	
5	30,000	1,084	$54,750	$51	$4.06	12	Modeling Only
6	15,000	806	$30,446	$38	$3.89	10	
7	15,000	757	$28,442	$38	$4.79	8	
8	15,000	727	$26,678	$37	$4.72	8	Modeling & Contact Management
9	15,000	690	$28,064	$41	$4.10	10	
10	15,000	725	$27,225	$38	$5.07	7	

Fig. 24.1 Table depicting payback in months after adopting certain data science techniques

technology being employed. Once modeling was introduced, this time was dramatically reduced, to approximately 12 months. Deployment of modelling yielded the largest benefit simply because no analytics was undertaken prior to this activity. Future analytics work will yield benefits, but they won't be as significant since the reference point is now modelled activity as opposed to random activity. The introduction of the campaign management system reduce the timed to payback from 12 months to approximately 8 months. As the retailer uses this system over time, more intelligence will be gained, thereby providing further insight in how to best reduce this metric (payback time).

Key Takeaways

- The challenge for retailers is to collect data at the individual customer level.
- Loyalty programs have represented the best course of action in attempting to collect individual customer data.
- Even without the capture of individual customer level, analytics can still be executed to determine store strategies or product placement strategies.
- In one retailer that we worked with, they were able to use their credit card database to market a variety of insurance products.
- The use of a campaign management system can augment information in being able to provide additional insights on prior promotional activity and its impact to the consumer.

CHAPTER 25

Business-to-Business Analytics and an Example

An Untapped Sector

Much work in predictive analytics and data science has been primarily focused around the business to consumer sector (B2C). Certainly, predictive analytics solutions have been applied to the B2B sector but it pales in comparison to what has been applied and learned in the B2C sector. Yet opportunities to gain more learning and knowledge exist where the same data science techniques can be applied to business to business (B2B). But there are some essential differences between the two sectors. In almost all cases, large businesses are typically excluded from any data science/data science exercise as the real data science/data science opportunity is to improve the marketing effectiveness towards the small to medium-sized businesses. Another key difference is identifying the key contact or decision-maker within the organization. Since our marketing and data science efforts are directed at someone taking action, then the ability to take action on behalf of a given company can vary based on the contact person at that point in time. By exploring the key contact or decision-maker within the organization, we may want to better understand the specific characteristics and behaviors of the key contact, along with the organization characteristics. But first, what kind of data is available for analysis? Can we capture job position and job role? In one exercise for a B2B software company, we analyzed job roles and job titles in terms of their response to a marketing offer. Differences were indeed identified by job title and job role and their impact on response. Additionally, we were also able to identify the prior engagement activity of contacts and how that may have impacted response. For example, we were able to determine the relationship between marketing response and the number of times a given contact is promoted prior to the marketing offer. Obviously, more information about how these contacts behave within the organization would have been useful in our analysis. But this is the challenge, as B2B customers will only yield information to the software company that is relevant for

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_25

billing purposes (i.e. job title and job role of contact) and not how contacts behave internally, or who might be the key stakeholder in most purchase decisions. Nevertheless, the overall modeling approach in developing predictive analytics solutions is identical between B2B and B2C.

Yet, even without this enriched contact information, B2B information has a significant advantage over B2C information in terms of available external information. In acquisition programs, external information is vital to the success of any analytics solutions. Within B2C exercises, most external information, particularly in the Canadian market, is at an aggregate geographic level in terms of where people live. In other words, the demographic information appended to individuals essentially consists of averages, percentages, and median values at a geographic level. The notion here is that consumers in that specific geographic area comprise the same demographic profile. Previous content in the book discussed the use of Stats Can data as the key data source for geo-demographic data.

Meanwhile, within the B2B sector, external data is at the individual company level, which is a clear and distinct advantage in developing acquisition targeting tools for B2B marketers. Often referred to as firmographic information, variables pertaining to employee size, tenure of the company, sales, industry category, etc. are collected at the individual business level. With this powerful data, it is not unusual to see the targeting of B2B prospects yielding lift results of above 500 between the top decile and bottom decile, which is clearly superior to what is traditionally seen in B2C acquisition programs.

Once a business becomes a customer, the existing data science approach in dealing with existing business customers is similar to what occurs for individual consumers. Type of customer, how long they have been a customer, and where they are located, all represent demographics that are similar to what we observe with regular consumers. Meanwhile the spending behavior and campaign/contact behavior represent the rich and robust behavior that will typically be used to predict the future behavior of that organization. Cross-sell, upsell, and retention models are all developed for existing business customers, not unlike what is observed on customers who are individuals. As mentioned above, businesses comprise three main categories based on size (large, medium, and small) which marketers consider first before developing any marketing plan or strategy. Indeed, it is the small business category where predictive analytics solutions offer the most significant benefits due to the huge volumes inherent in this sector. Yet, even within this sector, there is a sector of sole self-employed individuals who are quite unique from other small businesses. Although these individuals are mainly self-employed (i.e. one-person shops), one could argue that these businesses could be considered more like consumers. The same mechanics in understanding consumer individual behaviors and demographics is just as critical to understanding these self-employed entrepreneurs, since more emphasis will be placed on their individual behaviors and information rather than the company as a whole in trying to predict future behavior. Some practical examples from our experience are in the insurance and health sectors.

In the insurance sector, tools can be developed to target brokers who are most likely to sell a given insurance company's product. However, the challenge in developing these tools is the source data that is available in creating the analytical file. The information collected on brokers by many insurance companies tends to be sparse and inconsistent as broker data is collected from multiple sources. Another similar example of this type of challenge are sales persons/agents where, once again, sales agents can work for multiple organizations and data is again being collected in multiple places. The data challenge lies in trying to pool broker or sales data across multiple databases to create that one unified view of the broker or salesperson. It is this unified view of the sales person or broker that allows us to develop strategies for both high performers or high potential performers.

Within the health sector, rich individual-level data for doctors exists through organizations such as IMS or IQVIA. Examples of this doctor level data include type of specialization, university attended, tenure as a doctor, script behavior, etc. This can be extremely valuable to pharmaceutical companies as they can now develop predictive models which identify doctors who are most likely to prescribe a certain drug. Based on the type of solution, resources can then be better allocated to deliver higher return on investment (ROI) on a particular marketing/sales initiative.

Another challenge in dealing with businesses is the relative treatment of head offices and branch offices. For instance, should branch offices be treated as separate marketing entities as opposed to head offices only? This may not be totally relevant for small businesses, since most of them would only have one office. Nevertheless, there will be some small businesses that have multiple offices where again a true understanding of the specific marketing objectives will dictate the appropriate level of treatment.

The process of developing B2B predictive models is no different than that applied to B2C models. The significant difference once again is the domain knowledge in dealing with data that is unique to businesses such as firmographic information and self-employed entrepreneurs such as doctors, insurance brokers, and sales agents. The literature in predictive analytics often deals with consumers, whereas the development of models in the B2B environment might not garner much attention. This scenario limits our ability to gain knowledge as practitioners of data science since there is a whole body of work and practices that have occurred within this sector. As practitioners, we should encourage the sharing of this work and learning. Practical experience in the B2B sector does provide unique levels of experience and, ultimately, learning beyond the B2C sector, but which would be invaluable for those practitioners who are making the leap from the B2C world to the B2B world.

One example case study here is a small or medium-sized business rather than an individual. The company in question, a courier company, wanted to refocus its marketing efforts and essentially initiate a CRM program. As discussed previously in this book, management wanted to identify the company's best customers and profile them in order to better understand them. Yet, in creating or

establishing metrics to identify the best customers, only purchase behavior was considered in order to keep things simple. However, this seemingly simple metric was not so easy to determine. Purchase behavior was often very sporadic and infrequent. The first challenge, then, was to determine the appropriate time frame for a small to medium-sized company in purchasing this company's courier services. In other words, do active customers make regular purchases in intervals of 1 month, 3 months, 6 months, etc.? Once this time frame was determined, the company could then use the purchase metric as a proxy for value, albeit that this was a simplistic measure.

A further component to the analytical exercise—in addition to segmenting customers based on value to identify the best customers—was segmenting customers based on recent behavior. Were customers increasing or decreasing their purchase patterns, or were they perhaps defecting? For segmentation based on both value and behavior we adopted the VBS approach, or Value/Behavior Segmentation. The first challenge, though, was to determine the appropriate time window for purchase behaviors. The approach here was to consider purchase patterns within different time frames to see if these patterns were stable or volatile. These windows were then split equally into "before and after" periods to capture changing behavior in the periods before and after the time window examined. For example, with a one-month purchase window, analysts would look at purchase activity in one month and compare it to activity in the following month. Similarly, with a three-month window, analysts would examine activity within a three-month period and then compare this to the activity to the subsequent three-month period. Several options were considered in defining the purchase window. Several purchase window options were examined:

- one month purchase
- two months purchase
- three months purchase

Then customers were categorized into the following five groups based on these purchase patterns:

- Growers, that is, customers who were in the top 10% of purchase change behavior based on purchase amount and volume in the before and after time periods
- Decliners, that is, customers in the bottom 10% of purchase change behavior based on purchase amount and volume in the before and after time periods
- Stable, that is, customers not in the first two groups but with activity in both the before and after time periods
- Reactivations, that is, customers with no activity in the before period but activity in the after period

- Defectors, that is, customers with activity in the before period but no activity in the after period

Figure 25.1 was produced to present our results to the business people.

This report looks at the three different options (one month as the purchase window, two months as the purchase window, or three months as the purchase window) over a 12-month period and demonstrates how the time windows were evaluated. This report considers only 10 records for the sake of simplicity; in actuality, hundreds of records were assessed. Although this illustration visually identified the appropriate purchase time window, statistics were also employed to determine when the segment definition becomes constant (i.e. variance analysis) as the time period is increased.

In this report, the one-month option demonstrates the appropriate segment for the given customer (small or medium-sized business). Extending the purchase window to two months, the segment definitions change quite a bit for the same customer and continue to change when the three-month purchase window option is used. With the four-month and five-month options (not shown here) for the same records, the segment definitions become more stable and consistent with what the three-month option showed. The implication here is that the appropriate window is three months, as the segment definitions do not change when longer purchase periods are considered. Accordingly, all the behavioral segments were determined using three months as the normal purchase window for this company's customers.

The value segmentation was the next part of this exercise. The courier's B2B customers were segmented based on purchase volume within the last 12 months. The chart provided in Fig. 25.2 shows the rank order of these active B2B customers based on annual purchase revenue. As the chart shows, the top 20% of customers make up the high-value segment and account for average annual sales of $770, making up 68% of all sales by that courier company. In contrast, the low-value segment comprises most of the customer base (55%) and has average annual sales of $25, contributing only 7% of all the sales by that courier company.

	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Option 1 1 month	Option 2 2 months	Option 3 3 months
Record1	$100	$0	$0	$200	$0	$0	inactive	defector	grower
Record2	$300	$0	$200	$150	$0	$75	reactivator	decliner	decliner
Record3	$500	$400	$300	$0	$0	$0	inactive	defector	defector
Record4	$0	$200	$0	$0	$50	$0	defector	reactivator	decliner
Record5	$0	$0	$0	$50	$75	$100	stable	grower	reactivator
Record6	$700	$600	$500	$400	$300	$200	stable	decliner	decliner
Record7	$0	$0	$400	$0	$300	$0	defector	decliner	stable
Record8	$0	$600	$0	$400	$0	$500	reactivator	stable	stable
Record9	$0	$0	$200	$0	$0	$400	reactivator	grower	stable
Record10	$450	$300	$250	$150	$100	$0	defector	decliner	decliner

*Determined that 3 months was the optimum purchase window

Fig. 25.1 Examining different pre- and post-purchase window options to determine appropriate purchase window

Fig. 25.2 Value segmentation for B2B

	Average Annual Sales	% of all sales captured by segment
High Value (Top 20%)	$770	68%
Medium Value (20%-45%)	$230	25%
Low Value (45%-100%)	$25	7%

Segment	$ value	Migration Rate	$ opportunity
High Value Grower	$800	1%	8
Medium Value Grower	$350	10%	35

Fig. 25.3 Example of $ opportunity for two segments in B2B

After identification of the value and behavioral segments, the dollar opportunity for each segment was calculated by multiplying the migration rate by the dollar value of that segment. Using the business knowledge of the courier company, we were able to infer a migration rate for each segment, and this allowed us to estimate the dollar opportunity for each value/behavioral segment, as is shown in the example of two segments in Fig. 25.3. This illustrates how this dollar opportunity was calculated for two segments.

Based on these two segments, medium-value growers represented a better marketing opportunity than high-value growers. This approach, once adopted for all segments, allowed us to rank-order all value behavior segments based on this dollar opportunity, and this yielded insight on how best to allocate the marketing budget.

Key Takeaways

- Data science in B2B presents very different data challenges than B2C, yet the approach in building solutions is identical.
- Acquisition exercises can present better solutions in that external data sources yield data at the specific company level, in contrast to the B2C exercise, where much of the information about consumers relates to the geographical area of their residence.

- One significant challenge is to identify the key contact person within an organization, which is often not directly available from the data.
- Simple, but pragmatic segmentation approaches such as VBS (Value Behaviour Segmentation) have been used in our previous work with B2B clients.

CHAPTER 26

Financial Institution Case Study

Background

A financial institution used business rules to target regular credit card customers to become gold card customers. Specifically, they used tenure and region of the country as the key segmentation criterion in selecting names. The results of using these criteria as list selection rules revealed that the targeted names barely outperformed the random lists. More importantly, the cost per new gold card acquired continued to increase from campaign to campaign. This condition was particularly worsened by the business objective of significantly increasing the gold card base.

Given these above facts, the organization required a solution which needed both to optimize the volume of gold cards and to do it in a cost-effective manner. The financial institution identified the need for a tool such as a predictive model since this would have the following advantages over the existing list selection criteria:

- Could be applied to names beyond the boundaries of the list selection criteria, such as tenure and region of country.
- Could look at variables and other characteristics beyond tenure and region of country that could achieve the objective of optimizing response rate.

The challenge, then, was to develop a predictive model that optimized response rate.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_26

Data Challenge

Ideally, the best environment for developing a predictive model would be a random sample that had been received promotions in a previous campaign. In this situation, a campaign which occurred several months earlier had a random sample of 40,000 names. This was mailed to a group of names which had spent over $1200 a year, or an average of $100 per month. The response information to this campaign was captured, with the response vehicle (application) containing the account of the regular card. With the account number, this information could be matched back to the marketing file in order to determine responders and non-responders. Both non-responders and responders were then matched to the following files by account number to append the following information:

- Name and address file containing tombstone-type information such as age, income, household size, etc.
- Credit history containing spending and delinquency-type information over a given period of time.
- Banking file containing the customer relationships, such as loans, deposit/checking accounts, RRSP's, etc.

As previously discussed, when building predictive models, the analytical file should be crafted into a pre- and a post- window. If the campaign dropped in July, for example, the only information after that month that is relevant for model building is whether or not the customer has responded. All other information, which is considered as potential model variables, must be viewed prior to July.

There were no challenges in matching or linking files together, since the matching logic used an account ID. Yet the real challenge was the richness of data and creating all the potential model variables. There was no data mart or marketing warehouse at the time of the model-building exercise. All of the information was available on legacy systems, implying that the data was contained in raw transaction and billing systems. The task of organizing and summarizing the data into meaningful information was the most time-consuming stage of this project as 300+ variables were created in this process. In addition, Stats Can data of more than 100 variables was appended to these records by postal code. By the end of this process, the analytical file had 40,000 records, with the dependent variable being response, and over 400 variables would be analyzed as potential model predictor variables.

Using the modeling methodology previously discussed, a solution was obtained which contained the following profile of a gold card responder:

- Higher behavior score
- Less likely to have an RRSP

- Likes to travel
- Does not live in Quebec

In addition to the targeting capability, which will be discussed below through the gains/decile reports, the profile variables provided insight on various communication approaches (less likely to have RRSP and like to travel).

Using a cutoff of 40% from the gains/decile reports as the cutoff, or 200M names, the model yielded an expected benefit of $96000 during model development, which represented the opportunity cost of promoting additional names without modelling to achieve the same level of responders with modelling.

The following test matrix was set up to test not only the model's effectiveness but also different communication pieces. You will observe that most of the names being selected are in the top two quintiles, along with the test communication, as the expectation is that these cells will deliver the best performance (Fig. 26.1).

The actual results of this campaign began to be felt quite early in the campaign due to increased traffic in the operations department as they had to process an increasing number of applications. The final results of this campaign were as shown in Fig. 26.2.

Fig. 26.1 Marketing matrix after implementation of model

% of File (Ranked Model Score)	# of Names Mailed	Test Cell	Control Cell
0-20%	100,000	1 - 95,000	1- 5,000
20-40%	100,000	2- 95,000	2- 5,000
40-60%	100,000	3- 5,000	3- 5,000
60-80%	100,000	4 - 5,000	4- 5,000
80-100%	100,000	5- 5,000	5- 5,000

Fig. 26.2 Actual test results

Campaign Test Results	Actual Response Rate
Modelled Control Names (@ 40% cut-off)	1.87%
Random Control Names (Combine all 5 cells)	1.10%
Actual $ Benefits of Model due to Saved Mailing Costs (@ $0.80 per mail piece)	$112,000

Key Takeaways

- This case study for a financial institution is a classic case of the 4-stage data science process. In this case, however, the solution was the development of a predictive model.
- Creating the analytical file with more than 400 variables was the most onerous and time-consuming part of the process.
- This one simple campaign yielded $112000 in benefits, which led the bank to develop data science as a core discipline of its business.

CHAPTER 27

Using Marketing Analytics in the Travel/ Entertainment Industry

The growing importance of analytics and data science has now evolved towards looking at solutions in other sectors, such as travel and entertainment. Let's look at sports as one example of entertainment. Here, a flagship team like the Toronto Maple Leafs or New York Yankees does not need to be very effective in achieving the goal of putting more fans in stadium seats in a cost-effective manner. The reality for these organizations is that the product they offer is highly price-inelastic, meaning that the price charged for the product can easily be increased or decreased with only minimal impact on overall ticket demand. The use of analytics and data science within this context is meaningless.

However, for most sports organizations, putting more fans into seats in a cost-effective manner is a primary objective. Historically, sports marketers have attempted to increase attendance and grow ticket revenue by simply allocating more marketing dollars to mass advertising. This, however, has not always achieved the objective of increasing ticket sales. Sports marketers need to understand the concept that all fans are not created equal, which is very similar to the standard CRM slogan of "not all customers are created equal". The capability to do this is based on data, and whether or not information is being captured and stored somewhere. Unless someone at the receiving end of a ticket purchase transaction is capturing information about the ticket buyer, the activity or event becomes unavailable for any future analytics. Historically, it was extremely difficult to capture this information since a significant portion of a team's ticket revenues might arise from unrecorded events such as the purchase of tickets with cash at an arena or a ball park. Today, in the area of e-commerce, this has changed: fans can digitally purchase tickets through a variety of different ticket platforms. In fact, the actual non-digital purchase of tickets at an event now represents the exception, with most organizations deriving most of their ticket sales online.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_27

This individual behavior of fans can now be used by marketers to expand information capture beyond just the initial ticket purchase to include concessions and merchandise. Through the increased ability to capture data at the individual level, marketers can develop better programs to encourage incremental spend in both ticket- and non-ticket-type transactions. This growing trend of online purchases has accelerated the marketer's ability to capture individual-level data and, more importantly, to use this data and information as a way to differentiate between customers.

Having considered sports, let's now look at the hotel sector, which is another significant facet of the travel/entertainment industry. Historically speaking, the needs of a hotel customer were relatively narrow, with customer mobility being a key limiting barrier. In those days, the population was arguably more homogenous, with many of the needs being similar. In today's more complex world, customer needs are much more heterogeneous. Making this challenge even more daunting is the fact that our constantly busy society results in less time to explore these varying customer needs. Yet, even with this reduced time, consumer expectations are not diminishing; they are, in fact, increasing as technology has provided the basic capability of "doing more with less".

Within the travel industry, customers have always considered their time at a hotel as an experience rather than just a visit. Activities such as fine dining, nightly entertainment, spas, and corporate seminars/meetings nurture this concept of "customer experience". It is easy to see that such a range of activities is going to have varying levels of appeal amongst a given clientele. Obviously, the role of data science and analytics can be quite significant in helping marketers to better understand these varying client needs.

The first task might be to conduct a basic customer value exercise in order to ultimately identify the best prospects. But, as with many analytical exercises, the concept of seasonality needs to be considered here, something which is arguably even more significant within the hotel industry. For many projects, marketers make the assumption that the rank ordering of customers is unaffected by seasonality. In other words, customers may spend more during different times of the year, but their relative spend to each other remains unchanged. Many analysts could certainly debate this notion, depending upon the specific industry. Most analysts would agree, however, that in the case of the travel industry, the issue of seasonality can potentially have a significant impact on travel behaviour. For example, the first person may spend $1000 annually as a casual traveler throughout the year, and they might be considered an average customer. Meanwhile, a second person spends $1000 annually, but this all goes on a tennis package for one week in August. Should this second person be considered the same as the first person since they had the same annual spend? Obviously, the answer is "no". Both people spend the same amount, but they are, in fact, very different types of customers. This notion of seasonality is significant when conducting any analytics exercise, particularly if consideration is given to the many hotels who will offer tennis and golf packages in the summer and ski packages in the winter. In addition to the issue of seasonality, there are

various services that may have more appeal to certain groups of customers. Fine dining and theatre may appeal to one group of customers, while spas and perhaps valet-type services appeal to another group. With varying interests amongst travel clientele, a cluster-type segmentation exercise would be very useful in trying to assess different needs to different customer segments. Experts with domain knowledge in the travel industry would certainly understand that there are distinct or homogenous customer segments. Using the data being captured on these customers, it is now possible to apply some science to identify what indeed are the truly distinct customer segments. Specific marketing programs and initiatives can then be designed against each of these customer segments.

Key Takeaways

- Big Data has resulted in the use of many different digital platforms with the ability to capture individual fan behavior.
- In addition to selling more tickets cost-effectively, organizations now can sell merchandise based on this individual behavior.
- The hotel industry can utilize their individual guest behavior in attempting to create unique customer segments with differing needs which fluctuate according to the different seasons.

CHAPTER 28

Data Science for Customer Loyalty: A Perspective

What truly constitutes customer loyalty has been an ongoing debate amongst members of the marketing community. This is concerned not so much with the semantic definition of customer loyalty, since most marketers would agree that this behavior is represented by a strong affinity or attachment to a given company's products or services. The differences arise when marketers attempt to measure or evaluate customer loyalty.

The implications here are that marketers need first to *define* metrics and measures of customer loyalty. Once established, it must then be determined if the data is readily available in the current information environment for the creation of these measures.

Is RFM an Accurate Measure of Customer Loyalty?

The most common approach taken to measure customer loyalty is to consider prior purchase behavior. One very common, quick, and pragmatic way of doing this is through the development of an RFM measure, where:

R = recency of last purchase

F = frequency of purchase within a given period of time, and

M = monetary value of purchase(s).

By combining these measures, RFM indexes or scores can then be produced to generate a relative measure of loyalty for each customer. The higher the score, the greater the relative loyalty.

Many loyalty experts and pundits will debate this idea as being overly simplistic and incapable of capturing the true essence of customer loyalty. They argue that there are groups of customers with high RFM scores who are not really loyal and who will happily switch companies or brands if they see a better offer. Their so-called loyalty is not really due to any attachment or affinity to the company or brand but because switching to another organization for the

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_28

same products and services simply involves too much effort and/or cost. This actual weak affinity may quickly be broken if this type of customer is presented with an easy, risk-free offer.

But should it matter that some high RFM customers are not really that loyal? Does loyalty really matter if these customers represent (in most cases) the most profitable customers? Usually, the value that these customers deliver to the organization will result in the creation of "Best Customer" marketing programs which target this group, regardless of the relative strength of their loyalty. Nevertheless, marketers would still want to identify their most loyal customers who are not currently their best customers. This means that we need to consider other, non-purchase behavior measures that can capture customer loyalty.

Using Marketing Programs to Boost Customer Loyalty in the Short Term

Whether the group is high RFM or high value (or both), the next challenge is how to acquire basic learning about the effects of marketing activities that boost the customer's attachment/affinity to an organization's products/services in order to ultimately increase a customer's loyalty/value. One way of evaluating the effects of these programs is to examine customer behaviour that is truly *incremental* to past performance.

Changes in customer behavior can be defined in three ways:

- Lift
- Shift
- Retention

Lift behavior can be described as increased usage of a product or service; *Shift* behavior represents the acquisition of new customers to a product or service; finally, *Retention* behavior involves maintaining a customer's current activity level with a given product or service. In each case, the behavior is evaluated to determine if it is truly incremental as a result of a specific marketing activity. The key in identifying these behaviors (*Lift*, *Shift*, and *Retention*) as being incremental is through the creation of an appropriate marketing program testing strategy. Specific groups of customers are split randomly into test cells and control cells, where test cells represent the group being exposed to the marketing activity and control cells representing the group not exposed to the marketing activity. After the execution of the marketing activity, difference in performance between both groups can be compared and, ultimately, determines the incremental created by the marketing stimulus.

The use of RFM enables the ability to identify groups that may have a better return on investment (ROI) potential. The use of marketing activities against highly profitable groups allows for a proactive impact on the behavior of

specific groups of customers. Successful marketing program testing allows us to determine definitively the incremental changes in customer behavior caused by marketing programs.

Defining Customer Loyalty over the Long Term

Whereas previous purchase behavior may be an adequate measure of customer loyalty in the short term, there are other non-purchase behavioral indices that must be considered when creating measures of long-term loyalty. Three customer behaviors that might be considered as key information in identifying longer-term loyalty include:

- Overall Marketing Response
- Customer Inquiries
- Customer Complaints

Overall Marketing Response represents the customer's total number of responses to all the marketing campaigns of a given company. A customer's ongoing desire to respond to campaigns over time could certainly be construed as a measure representative of a customer's level of engagement with the company. These responses could be related to both purchase and non-purchase activities.

Customer Inquiries generally represent a positive interaction as they reflect the customer's desire to obtain more information from the company. In this activity there is an implied sense of trust between the customer and the company. This trust factor can only grow if the outcome of each inquiry is positive in the mind of the customer. Conversely, *Customer Complaints* represent a negative interaction between the customer and company. The erosion of trust between the customer and company may occur with each complaint.

Database and digital marketers want to identify these traits or characteristics from information recorded on a database. Marketing Response can be captured if there is some kind of campaign management system within the organization. With this type of system, campaign information and responses to those campaigns are captured at an individual customer level, enabling marketers to identify the overall marketing response history for each customer. Another rich source of customer inquiry information is from a company's website. By examining web logs, it is possible to identify the number of page views of customers. The page views, in most cases, represent basic inquiries for information. If customers are exhibiting more page views within a given company's website, they are generating more inquiries on the products, services, and other information about the company, and they arguably have a more heightened interest about the company. In the telemarketing world, both outbound and inbound calls deal with customers who often inquire about other products and services, as well as expressing complaints.

How is this information organized into meaningful loyalty measures? Campaign management systems and their resulting tables are well organized into structured rows and columns. Information arising from the web is less organized in that the information is semi-structured or object-oriented. In both cases, this information can be summarized to obtain specific views of customer activity or engagement. However, in the case of a telephone conversation, the information of interest is within the dialogue between the sales or customer service rep and the customer and is essentially contained in an unstructured format. Unstructured data means that it is impossible to point to one specific field; instead, the entire text of the dialogue is reviewed to extract meaningful information. The concept of extracting meaningful data from unstructured data has become a primary focus of data science and, in particular, artificial intelligence (AI). As we will see later in this book, AI has been used in many ways when analyzing text data; the following chapter actually begins the discussion on text analytics. Through this discipline, conversations, comments, and anything that is of a textual nature can now be analyzed to extract meaningful information. In the case of extracting meaningful measures for customer loyalty, text analytics can be used to identify customer complaints or customer interest levels based on the conversations or comments between a customer and a sales or customer service representative.

The use of campaign management systems, the internet, and text analytics provide new ways of creating customer loyalty measures beyond basic, short-term-focused purchase behavior. Taking a longer-term perspective to the development of customer loyalty measurement ensures that broader customer engagement and attachment behaviors are incorporated into these measures. Marketers may then be equipped with tools they can use in executing initiatives that grow customer engagement and drive long-term customer profitability.

Key Takeaways

- The creation of metrics to measure loyalty can be a challenge as our real objective is to measure customer engagement.
- Customer engagement can be a nebulous term as simple metrics like RFM do not necessarily capture the essence of loyalty.
- The internet has provided the ability to observe one's browsing behavior despite the fact that there is no related purchasing.
- The advent of artificial intelligence now provides us with better capabilities in analyzing text data as another measure of customer engagement.

CHAPTER 29

The Data Discovery Process

The data discovery process is new to Big Data, but it has become increasingly relevant given the exponential growth of data in our increasingly digitized world. The purpose of the data discovery is to:

- Identify gaps/opportunities in the current data collection and management practices that will help improve a given organization's ability to leverage its data.
- Gain insights into the usage of current products and services and identify opportunities for the growth of these products and services.
- Recommend a plan that will help the company harness the full power of its data assets.
- Set out clear next steps that will help the company achieve the recommended plan.

Listed below is a 4-step process.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_29

Preparation

This step consists primarily of a series of meetings between key stakeholders at the company and the data scientist in which the company provides the data scientist with more in-depth information on the current business, the current data environment, and the current data-driven initiatives. The areas that we address are:

- Understanding business issues and challenges
- Reviewing existing company data-driven products and services
- Understanding the number and type of data sets available e.g.
  - Customer data
  - Transaction data
  - Campaign data: email, social media, direct mail, telemarketing, etc.
  - Website data
  - Social media data
  - External/open source data sets that are used by the company
- Overview of the size and scope of each data set
- Understanding of current measurement and reporting practices

During this phase, the data scientist will also review any relevant documentation supplied by the company to impart a better understanding of the business needs, e.g. sample reports.

Based on the above information, the data scientist will finalize requirements for the data audit, including agreement on which data sets and how many years worth of data will be included in the data audit.

Data Audit

The next step is to conduct a data audit, which is the process discussed in Chap. 5 of this book. As discussed in that chapter, the purpose of this exercise is to examine the following:

- Quality of data
- Data integrity
- Usefulness of the data for data analysis

With Big Data, the data audit is extended to exploring a much richer and vast data environment arising from websites, social media sources, external public data such as Statistics Canada, or other publicly available open data sources that would be relevant for the company and project.

Once we obtain the insights andd learning from this process, we can then determine how to link all the data and to create and determine the appropriate features/variables that are relevent for the given project.

Preliminary Analysis

To better understand company's data and potential opportunities going forward, one can perform some basic analysis on the data reviewed in the audit. Potential areas for analysis include:

- Snapshot view of the overall customer base
- Segmentation of customers by:
  - Value segmentation
  - RFM (Recency of activity, Frequecy of activity, Monetary value activitiy)
  - Value–Behavior-based segmentation
  - Cluster segmentation
- Profiling of segments
  - Conduct analysis to identify what customers look like within a given profile segment.
- Customer migration reports
  - Insights on how customers migrate between segments based on value and behavior

Much of this analysis above has already been discussed, with examples throughout this book. But another good example is a customer migration report, such as that listed below.

	Current Year Activity				
Value Segment Previous Year	Growth	Stable	Decline	Inactive	Total
Best	---	41%	53%	7%	36,812
High	8%	23%	44%	24%	73,622
Medium	9%	21%	14%	56%	147,240
Low	14%	11%	---	75%	110,430
Total	**9%**	**20%**	**20%**	**50%**	**368,104**

In this above example from an organization with a loyalty program, we observe migration patterns that are clearly disheartening to the company. 7% of their best customers and 24% of their high-value customers become inactive in

the next year, yet they account for close to 75% of their overall business, as seen in other reports. The decline and inactive behavior in both the best and high-value segments simply reinforces the notion that their better customers are becoming more disengaged, which would suggest a marketing strategy geared more towards cultivation and engagement of customers.

This illustration simply reinforces the real purpose of this preliminary analysis, the intention of which is to identify areas for deeper analysis and make recommendations on how the data can be leveraged going forward.

Recommendations

A summarization of findings and insights is then produced into a detailed report, including short-term data strategies. Among the areas that may be included are:

- How to leverage data to create further customer engagement
- Data collection opportunities and goals
- Additional data analytics and/or data modeling initiatives that would further enhance the company's ability to leverage its current data resources
- New opportunities to leverage the client's relationships with other partners and suppliers
- New approaches to augment existing analytics and reporting capabilities which, of course, include the design and creation of the right measurement framework.

A prioritization of tasks and activities that could be undertaken in the next 6–12 months. See an example of report below.

Segment	# Customers	Avg. Value	Strategy	Projected RR	$ Growth Potential	Segment Potential
New	54,752	$80.40	Dedicated Welcome & Onboarding	15.00%	40%	$264,123.65
Best in Decline	9,755	$623.53	Retention	5.00%	20%	$60,825.35
Potential Best	22,086	$147.08	Up-sell	10.00%	30%	$97,452.27
Underdeveloped ProductX sales	15,203	$248.00	Cross-sell	2.50%	15%	$14,138.79

In this above report, the outcome of the analysis has identified the top four initiatives based on segment potential. The numbers in the segment potential are derived by multiplying the # of customers × avg. value × projected RR (conversion rate) × $ growth potential. For example, the segment potential in the first row for New is 54,752 × $80.40 × 15% × 40% = $264,123.65.

The intention of this report is to establish prioritization and to look at the segment potential dollars in a relative rather than an absolute manner. In this above example, it is evident that marketing initiatives and activities should be focussed around new customers.

Although I have summarized the process in just a few pages, this type of exercise is very labor-intensive as much time is spent in both acquiring business understanding and also knowledge of the data environment in order to address the key business challenges. It is truly an exploration, or data discovery, process.

Key Takeaways

- The data discovery, which is a very strategic and open-ended exercise, can be structured into a four-stage process.
- At the end of this exercise, a set of recommendations are produced which address the key challenges identified in this exercise.
- One component of these recommendations is a priority report which outlines and prioritizes key initiativees that have been identified in this process.

CHAPTER 30

Analytics and Data Science Within Insurance

The primary advantage of using analytics and data science for insurance is the ability to more effectively create a rating structure or a price premium for a given policy. Historically, challenges existed on how to build the best tools in this area, particularly the development of those related to multivariate analysis (MVA). In other disciplines, such as marketing and credit card risk, the notion of using predictive analytics and any associated MVA tools had a much stronger history than its application within insurance. Although actuaries have made extensive use of mathematics for well over a hundred years, the ability to use data from a data science perspective is a relatively new phenomenon. To fully leverage the results of any MVA tool, hundreds of thousands of individual policy records, each with several hundred variables, need to be created, which is the core skillset of the data scientist. At the time that the first edition of this book was written, many insurance companies did not have the ability to really leverage the rich data environment of their insurance companies. Since then, the situation has changed significantly, with virtually all such businesses now establishing data science as a core competency. Insurance companies either hire data scientists directly, or hire actuaries who are trained in this discipline. As stated throughout this book, although math is important, it is the data itself, and what practitioners do with it, that provides the bigger impact in a given solution.

Newer technologies are now being employed, necessitating an increased need for data science skills. Tools such as telematic devices can now measure real-time driving behavior. Prior to the development of these tools, data scientists would use the information supplied by the driver reported on their application form. This was certainly better than nothing, but practitioners will always want to work with data that is observed rather than what is simply being reported. Similarly, in the health and life insurance sector, the use of so-called "smart wearables" can generate data related to cardiovascular and respiratory

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_30

behaviour while an individual is exercising. Some insurance companies are building their own customized large language models (LLMs) which can be applied when customers contact their call centers. As the world becomes increasingly digital, consumers are seeking out online platforms which provide pooled services, offering individuals information about the entire insurance market. This means that insurance agents and brokers are becoming increasingly irrelevant. These online platforms, and their digital nature, with its capacity for real-time, individually tailored responses, are providing a much richer data environment for the practitioner. All these developments simply further reinforce the significance of data science within insurance.

Artificial intelligence (AI), of course, is now commonly used in some manner by virtually all organizations. But again, there needs to be some oversight regarding the development and use of these tools within individual business settings. This requires an individual who has a basic understanding of how these solutions are built. To some degree, the person needs to understand the math and the parameters that can build different solutions, as well as how they can be applied. More about AI and its application will be discussed in the next chapter.

Key Takeaways

- The old actuarial approach of simply using math and statistics to create insurance pricing structures is no longer sufficient.
- Data science is now a core skillset of insurance companies.
- New technologies, such as telematics, "smart wearables", and online broker platforms all enrich the data environment of the data scientist.
- AI and the use of LLM models further reinforce the need for data science as a core competency of insurance companies.

CHAPTER 31

Artificial Intelligence

There has been much discussion, some of it almost euphoric, about artificial intelligence (AI) and how, in many ways, it will establish a new framework for the Information Economy of the 21st century. Let's be clear here from the outset. Much has been written about AI, yet its creation involves complex mathematics which very few people really understand. As a well-known colleague of mine recently observed: "Math is not a good subject for most people". Accordingly, he simply spends all his time communicating its benefit to the business as he and his fellow data scientists deal with the math. This is fine in terms of what my colleague is trying to achieve. But the current discussion, as we all are aware, goes beyond just the specific benefits to a given business. One need only look at a magazine or newspaper to witness the volume of content surrounding AI.

I have spoken at numerous seminars to try and demystify this topic not only to business leaders but also to their data scientists. It is, of course, easier to achieve this objective given the specialized nature of the conferences which range from AI, machine learning to predictive analytics. At a more public mainstream level, however, there is still misinformation about its use, and particularly its limitations. Often enough, it appears that these changes seem daunting and directing us towards an unforgiving economy where there are significant job losses across all industry sectors. Yet perhaps a deeper understanding of the technology allows us to identify the opportunities that may emerge from AI. Are there current limits to AI that may, in fact, represent new opportunities which currently do not exist?

Before delving more deeply into this topic, let us be clear what AI is *not*. In some conferences and articles, you will hear that AI encompasses any type of predictive models/machine learning algorithms. *This is a fallacy*. It is, in fact, much more specific in that true artificial intelligence is about deep learning,

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_31

which involves the use of the mathematics of neural nets where you have an input layer, multiple hidden layers, and an output layer.

Complex algorithms using different optimization approaches have been developed in delivering solutions which can demonstrate, and indeed have demonstrated, great performance over the past few years. But what has caused this tremendous upsurge in AI interest, given the research has been around for decades?

Like many processes, AI needs fuel, and in this case this consists of volumes of data. Technology has always been the limiting barrier in being able to process huge volumes. However, this is no longer the case as companies can now more fully leverage Big Data technology, and its parallel data-processing approach.

With all the discussion raging around AI and its use across many business disciplines, seasoned data science practitioners are often asked whether AI should be the predictive analytics or machine learning tool used in building these above solutions. Yet different disciplines will have different ways of evaluating success. The disciplines that have been using AI, and which have seen the greatest impact, are the areas of image recognition and NLP (natural language processing). These are the disciplines that are yielding the groundbreaking improvements in AI.

There are many textbooks on this subject which are filled with complex mathematical equations, and I would certainly be doing these esteemed authors a disservice to go into any more detail on its mathematical complexities. Yet a basic high-level understanding of the technology can enable someone to understand why this mathematics is referred to as artificial intelligence. The utilization of layers (input, hidden, and output) yield an architecture in which the machine checks a given solution and compares it to the actual outcome. If the difference is not within an acceptable range, the machine then adjusts the weights of its inputs, where the inputs are the data. These adjustments represent iterations where the machine is learning to optimize the solution and hence minimizing the difference between the predicted and the observed. In a way, this is comparable to how the human learns. Think of any sport where improvement only occurs after much practice. Continued practice reinforces those neurons to fire the right impulses to those required muscles and in effect reduce the error between the desired result and the actual result.

The fundamental question here is how extensive this artificial intelligence in its capacity is to replicate human intellectual capabilities. Are we at the crossroads yet where the human's intellectual range and depth can be replaced? In working within this field as a data scientist, my intuition is that the machine is nowhere near the intellectual capabilities, and this is best demonstrated by the chart below (Fig. 31.1).

In this chart, consider the lower part of the brain where AI has achieved its great accomplishments, image recognition and natural language processing. Note that these functions are considered lower-level human functions since they are located closer to the brainstem. The more cerebral thought processes

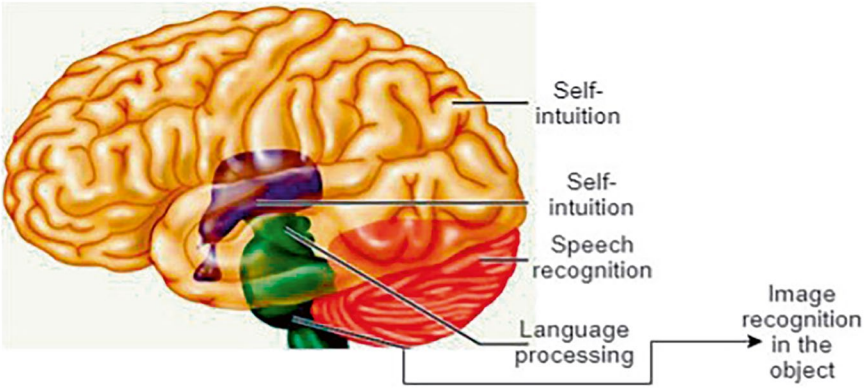


Fig. 31.1 Colored elements of brain represent the focus of AI activities

of the brain, which are located within the upper part of the brain and far from the brainstem, do not at this point indicate any advanced AI functionality. Why should this be the case?

If we consider the math again, the processing of AI in its current architecture moves very much in a linear type of process rather than laterally, as is the function of the higher-level type processing. Of course, this capability occurs in the higher areas of the brain well beyond the brainstem. However, the latest developments in generative artificial intelligence (GENAI) and in tools such as CHATGPT appear to demonstrate capabilities that are moving more in the lateral direction of the brain. Cognitive intuition or learning seems to be accelerating, demonstrating the increased thinking capacity of AI. The use of large language models (LLMs) is the key to providing these solutions. The basis for these solutions was discussed in the chapter on text mining and text analytics and the process we use in creating the actual solution.

For now, however, let's try and think of some good examples to highlight these key differences between the human and the machine. Nowhere is this best illustrated than in my discipline of data science.

The first example deals with problem solving. Let's say we have been asked to build a model to target new donors for a particular not-for-profit organization. Our domain knowledge of the not-for-profit business provides us with information that perhaps we should be exploring lapsed donors and reactivating these types of donors, which might be less costly than an acquisition donor program. But, of course, this is something that we would explore and then determine which is the best strategy. Another good example might be to optimize the acquisition of new customers for a telecommunications company. Our understanding of the business has indicated that retention is a real problem among new customers, thereby suggesting that our desired modeling objective is not only new customers but also new customers that will not defect in the

short term. In both cases, the brain has thought laterally in bringing together knowledge which can be used to better identify the problem.

A second example might involve feature engineering. In certain applications, such as image recognition and NLP, the machine and AI does a much better job of creating and manufacturing new features through the hidden layers which provide more value to the algorithm. Essentially, it automates the feature engineering process. However, my career experience has been in focusing on building predictive models to predict a desired consumer behavior. In this case, the creation of new features or derived variables in any modeling solution is arguably the most critical factor that a data scientist can bring to any solution. Let's look at some examples. Nowhere is this more prevalent than in the many cases where we must join files together, in which the relationship is a many-to-one relationship. It is in these situations that the data scientist must become intimate with the data to determine how to best summarize it in terms of creating ratio, penetration type variables and the basic stats regarding each variable, such as means, mins, maxes, standard deviations, and maybe even more complex stats such as log transformations. This work, although often underrated with regard to its value to a given project, provides the rationale as to why data scientists will spend 80%–90% of their time in creating new variables. In these consumer behavior cases, the machine is used to simply code the desired features as determined through lateral thinking by the data scientist.

Besides this type of variable creation mentioned above, variable creation will, in many cases, contain information that is more relevant to that company and its industry. Some great examples of this are to be found in the insurance business, where my domain knowledge continued to increase overtime as I was engaged in more projects within this sector. Here are a couple of examples. In predicting auto claim risk, it was readily apparent that age and years of license were always key variables. This was a clear case to create an interaction variable encompassing both variables. As I continued to acquire more knowledge about the business, we also determined that the event period in predicting risk should not always be based on the term length of the policy; rather, it should potentially be split into multiple time periods based on when a major change occurred within that term period. These changes, often referred to as mid-term changes, were identified as variables such as change in residence of the policyholder, occurrence of a claim, additions of new persons to the policy, etc.

In all these above examples, the question is whether an AI tool could have been able to automatically acquire this learning and build it into its solution without any human intervention. At this stage of its development, the answer is clearly "no" due to the limiting factor of linear-type processing. In contrast to the present-day example of the computer using AI, the human brain can leverage its lateral-type processing, thereby allowing the data science professional to develop the types of solution seen in these above examples.

In my early days as a data scientist, data mining publications (all the printed variety at that time) continued to publish research studies on the accuracy rates of various AI techniques. Many of these studies involved the use of AI to

improve the existing accuracy rates of image recognition, which were ranging in the neighborhood of approximately 45%. At that time, I realized that more advanced mathematics was being utilized than the typical mathematical tools I was using to predict marketing response and/or credit risk. However, I did understand that there was much work being done in feature engineering and the use of neural nets or deep learning to improve the image recognition rates.

Now fast-forward to circa 2012. Big Data, and our ability to process it at incomprehensible faster speeds when compared to the "tortoise-like" data-processing speeds of the 1990s allowed advanced analysts and data scientists to consume significantly larger volumes of data. Much like a car needs gas to run, AI and deep learning needs large volumes of data for its algorithms to build more optimal solutions. Data was the missing ingredient in the early studies of the 1990s. With its newfound fuel (DATA), the accuracy rates of AI techniques or deep learning in image recognition rose from 45% in the 1990s to well over 90% today.

As a data scientist, however, objectivity is the reality of our discipline. In other words, does AI or deep learning apply in all business scenarios? Can these incredible improvements in improved accuracy rates in image recognition translate to the world of predicting consumer behavior from both a marketing standpoint and a risk standpoint?

We continue to develop models using traditional techniques which our clients understand both in terms of performance and business benefit, but also, equally importantly, in being able to easily explain the key drivers of a model. Let's examine this more closely as we look at both issues of explainability and performance in assessing whether or not to implement a given model.

With regard to the first issue of explainability, using the more traditional techniques we could produce the following type of report (Fig. 31.2):

This report for a marketing response model can be clearly understood by nontechnical people as it highlights three key facts:

- The final model variables
- The variable's impact on response
- The variable's overall importance (contribution) within the model

Model Variable	Impact on Response	Contribution to Overall Equation
Behaviour Score	Positive	40%
Has an RRSP Product	Negative	30%
% of Credit Limit Used	Positive	20%
Live in Prairie Provinces	Negative	10%

Fig. 31.2 Example of final model variable report

Furthermore, the business stakeholder in this example could understand the profile of a responder and distinguish the strong drivers from the weak drivers of the model.

In the example of deep learning, where we have our input variables entering the neural net through the input layer, the hidden layers represent a form of feature engineering in that the weights of each input variable strive for optimization. Optimization routines can vary, but the more common routines utilize various versions of gradient descent which is essentially looking at the rate of change in attempting to converge to a solution that minimizes the difference between the predicted output and the observed output. Convergence occurs when this rate of change between two iterations is at a minimum or close to zero in the solution's attempt to minimize the overall error rate. To some degree, each hidden layer represents an improved set of features until we arrive at the final output layer, which represents the final solution or algorithm. But the final solution or algorithm represents a composite picture of input weights for each variable within each hidden layer. Given this scenario, it is easy to understand the complexity in trying to achieve a good understanding of the importance and impact of each variable on the desired modeled behavior. However, there have been some approaches that have attempted to measure the feature importance of a given variable. One such approach is to delete one variable at a time and to determine the change in score which is attributed to dropping this variable. This is done for each variable by looking at the overall change in score as a measure of importance. Now let's look at the second measure of model performance.

In much of the literature on AI and improved model performance, the discussion focusses on error rates between the predicted output and the observed output. Yet, in the world of consumer behavior at both the prospect and the customer level, the use of decile or gains tables represents a key business tool in assessing overall model performance. The chart (Lorenz curve) in Fig. 31.3depicts the modeled output of a response model when the model is applied against a validation sample.

Here, the *y*-axis represents the actual or observed response rate while the *x*-axis represents deciles where records are sorted by descending model score into 10 groups, with decile 1 having the highest predicted response scores and decile 10 having the lowest predicted response scores. Model performance is then determined by the steepness or slope of the line. A model with no performance yields a flat line as seen above which means that the model is delivering a random solution. Our objective in building models is to maximize the steepness of the line presuming that we have dealt with any potential overfitting issues.

In evaluating AI as a modeling option for consumer behavior, we want to observe whether or not AI can actually improve the slope of the model over and above traditional modeling techniques. We did some work in this area by looking at many different models and comparing traditional techniques to deep learning or AI models. Most of our research indicated virtually no

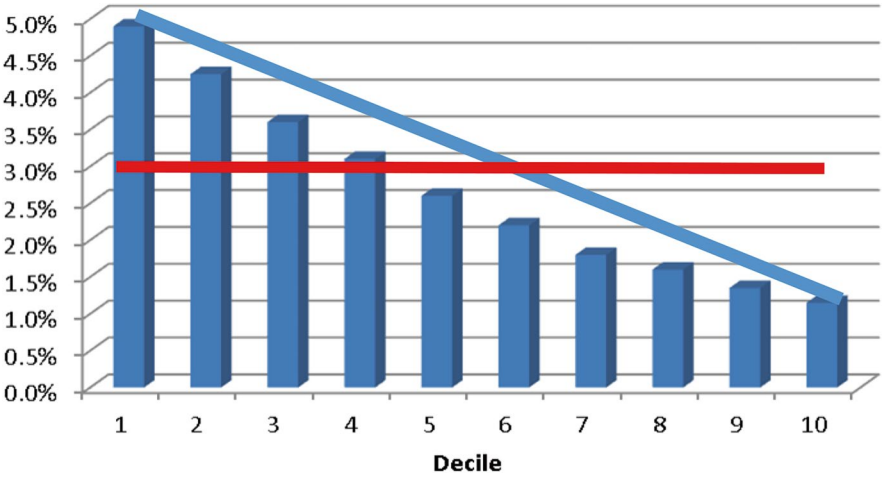


Fig. 31.3 Example of visual display of response model performance by model decile

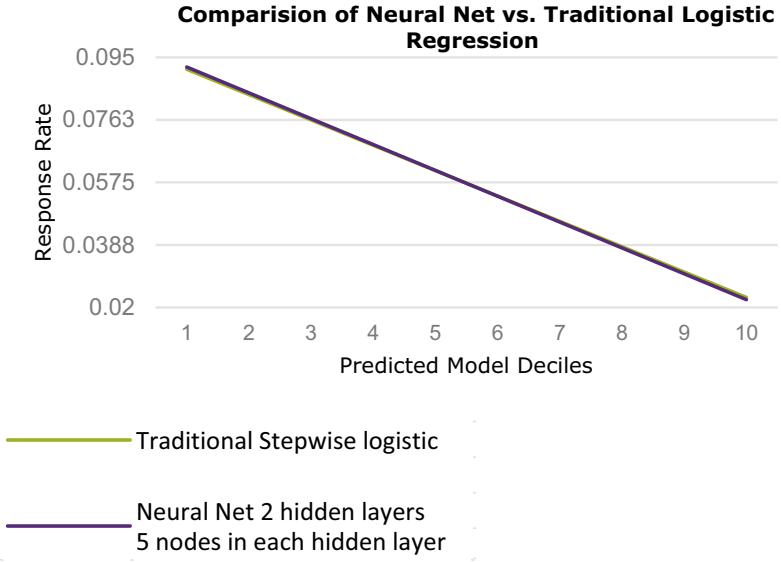


Fig. 31.4 Comparison of neural net vs. traditional logistic regression in small data environments

improvement in model performance as depicted by the steepness of the slope, which was the same for both traditional models versus deep learning. Rather than list all the models we utilized in this research, I have included one example (marketing response model) which is representative of much our findings in this area (Fig. 31.4).

In the above example, the lighter and bolder lines have virtually identical slopes indicating virtually no difference between AI (deep learning) and logistic regression in predicting marketing response.

Yet, our findings were not 100% conclusive, as seen by the chart (Fig. 31.5) which depicts the model's ability to predict property claim loss.

In this example, the neural net or deep learning model clearly outperforms the traditional stepwise logistic and would deliver increased business value in being able to make better decisions regarding the pricing of property policies. But why did we see such an improvement? First, we had access to much larger volumes of data than our typical modeling situation. Second, we also believed there was less random error in predicting a property claim loss than in trying to predict a human being's likelihood to respond. Besides having access to large volumes of data, the element of a strong signal-to-noise ratio is another critical factor in deep learning being able to deliver superior model performance.

Nevertheless, the ability to explain the neural net algorithm still looms as a challenge for many businesses in being able to embrace these solutions as part of their business processes. However, neural nets and deep learning are here to stay, despite their explainability challenges. But it is incumbent on the research community to devote more efforts to how data science practitioners can explain these types of solutions to their stakeholders. Much research is already being done in this area alongside efforts to utilize AI in a non-Big Data environment. As always, the data science community looks forward to the results of these efforts.

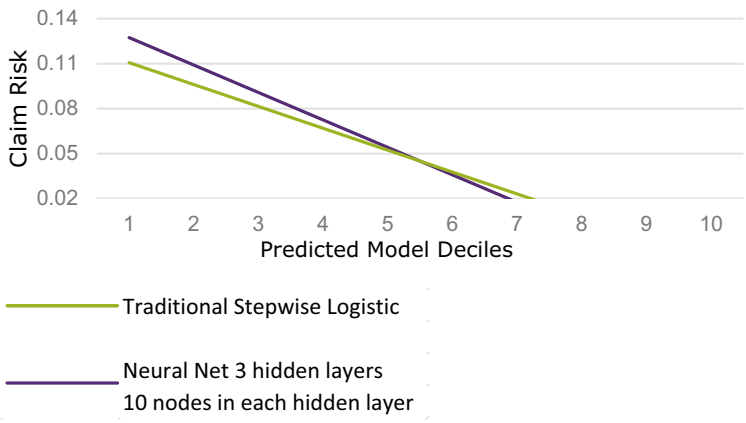


Fig. 31.5 Comparison of neural net vs. traditional logistic regression in large data environments

What the Data Scientist/Data Miner Needs to Know When Using AI

Two of the most familiar common-type neural net architectures are convolutional neural networks (CNN) and recurrent neural networks (RNN). For practitioners, it is important to know that CNN is used in image identification and classification while RNN is used for NLP and text recognition.

Much like the other modeling algorithms that practitioners use (multiple/logistic regression/, discriminant/SVM, decision trees, etc), the user acts like a pilot in understanding the "dials" or parameters associated with deep learning. For example, the most important component would be the features or variables that are a part of the input or analytical file. In consumer behavior models, much of the so-called derived variable process or "feature engineering" is manual. Here the data scientist utilizes both the creative thinking of the data scientist and the domain knowledge of the given business or association to create meaningful variables in the input file. This process is the most time-consuming part of the process, but it is one that delivers the greatest value in terms of model performance. Typically, it is not unusual to have 90%–95% derived variables, with the rest being source variables. This data science skillset of variable derivation and feature engineering has been emphasized throughout this book.

In the area of image recognition and NLP, the feature engineering component is automated. In both these applications, the pixel (image) and the text (NLP) are converted to vectors. These represent the input variables into a deep learning algorithm. The deep learning process, through its hidden layers, transforms these vectors into outputs, which are then used as inputs as the data moves to the next layer. This is all automated by the algorithm. Figure 31.6 shows the simple learning network as compared to the deep learning network which is what is really used in all image and NLP applications. In the deep learning example, we have three layers, but in both these applications (NLP

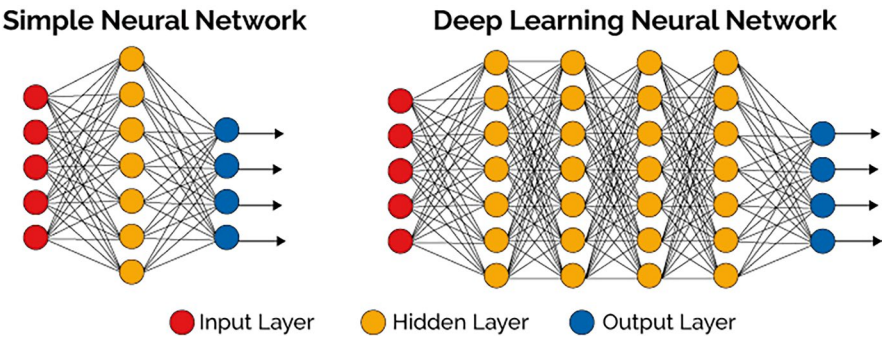


Fig. 31.6 Simple neural net architecture vs. deep learning neural net architecture

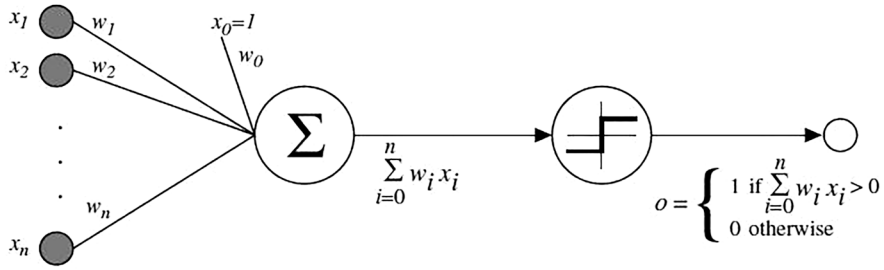


Fig. 31.7 Schematic of weights (w) and nodes (x)

and images), the number of hidden layers could be significantly higher than three.

Notice that within each layer, we also have several nodes in each layer which are connected, as depicted by the lines between each layer. In this network, the nodes themselves strive to improve their learning. Figure 31.7 looks at one diagram in which the node is referred to as the perceptron. There are weights (w) and inputs (x) which are the ingredients in activating these perceptrons or nodes.

In another diagram (Fig. 31.8), we schematically observe the mathematics behind the activation of these perceptrons.

It is these weights which are constantly adjusted to minimize the difference between what is observed and what is predicted. The concept works as follows:

1. Start off with a perceptron having random weights and a training set
2. For the inputs of an example in the training set, compute the Perceptron's output
3. If the output of the Perceptron does not match the output that is known to be correct for the example, perform the following:

    - If the output should have been 0 but was 1, decrease the weights that had an input of 1.
    - If the output should have been 1 but was 0, increase the weights that had an input of 1.

4. Go to the next example in the training set and repeat steps 2–4 until the Perceptron makes no more mistakes
5. The algorithm or solution is formed when the minimization of error between the observed result and the predicted result impact the weights in such a way that the change in weight is very minimal, hence reaching a point of convergence.
6. Of course, this solution to reduce overfitting needs to generalize well on unseen data such that the solution minimizes both the variance as well as the bias (i.e. generalization)

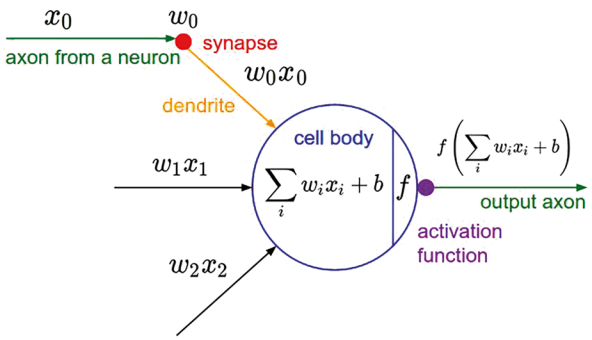


Fig. 31.8 Activation functions of nodes/weights and output

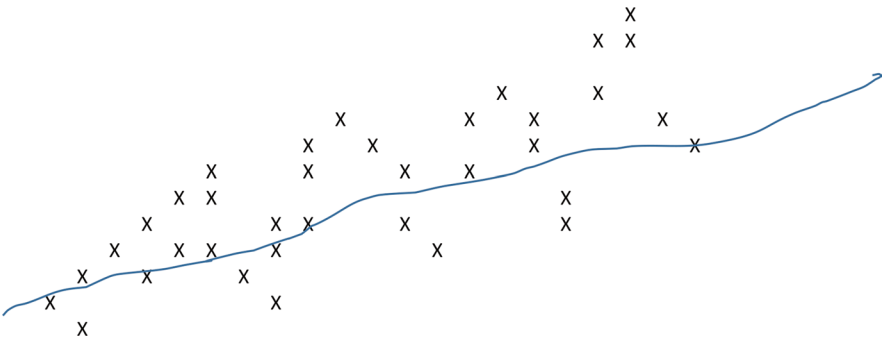


Fig. 31.9 Plotting a regression line through the points

So where is the non-automated process of deep learning and how can the data scientist add value to this process? A variety of user parameters provide flexibility to the data scientist. In the next few pages of this chapter, we discuss some of the parameters.

Once the input file is created, be it image, NLP, or consumer behavior, the user can then determine how to split the file into training, testing, and validation. This is a very important and critical process in minimizing error, which is a given for any modeling algorithm.

The Issue of Bias Error vs. Variance Error in AI Algorithms

Essentially, we are trying to arrive at a compromise in which we minimize error due to bias and the error due to variance. One good example of variance error is the one observed in the training dataset where, for example (see Fig. 31.9), we have a nonlinear distribution and we attempt to fit a linear regression as the best fit model.

In the above chart, the linear regression represents the fact that there is an upward trend, but it is clear that there is much variance error (training) here which is also referred to as underfitting.

Figure 31.10 offers a great example of overfitting which speaks to bias error as the model will not generalize on new unseen data.

As we can see here, the model is almost perfect as the algorithm depicted by the line captures virtually every X observed behavior on the line. It is almost certain that this pattern or trend depicted by the model will not be exhibited in new unseen data.

Essentially, what we truly require is a model that is a compromise which balances the minimization of error seen in overfitting (bias) against the error in underfitting (variance) (Fig. 31.11).

Here we still have a nonlinear type model algorithm, but one that best compromises the two errors of bias vs. variance.

Among the other user parameters are dropout rates, learning rates, and optimization and activation functions. Dropout rates help to improve the generalization capability of a neural net model by dropping out nodes across the network. The user can define which nodes to drop within each layer. A good dropout rate is one where the overall error seen in validation is minimized, recognizing that the error includes both bias and variance.

Learning rates are used within the overall optimization routines, where a high learning rate can cause a model not to converge to a saddle point within the distribution while a low learning rate can result in a model taking a very long time to reach its point of convergence i.e. minimum or maximum point on the saddle point (see Fig. 31.12).

Ultimately, we are looking for a point that is, again a compromise between a low and high learning rate that ultimately minimizes the error.

The activation function is another parameter that can be selected by the user. These functions comprise the following, as depicted in Fig. 31.13 (tanh, sigmoid, linear, and RELU-rectified linear unit). It is these functions that determine the predicted output of the solution.

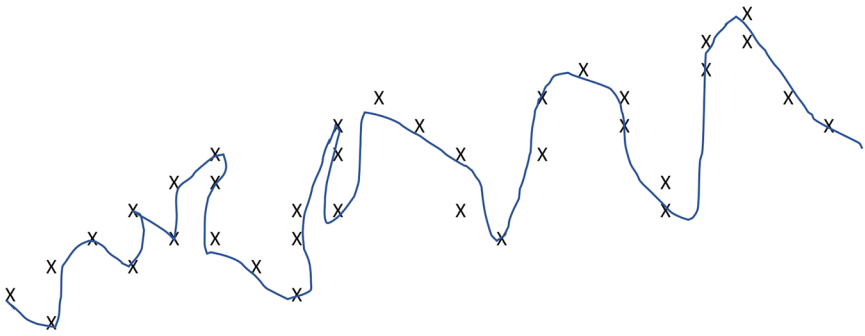


Fig. 31.10 Plotting a non-linear neural net model that overfits

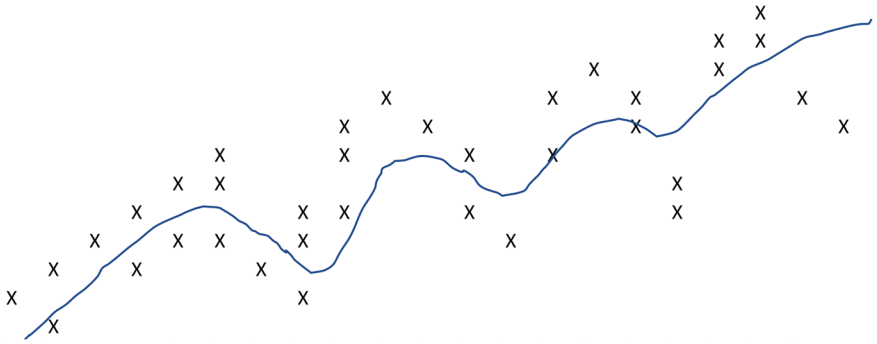
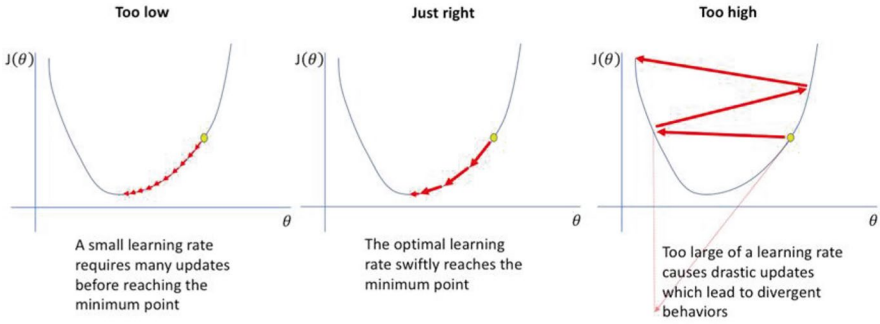


Fig. 31.11 Plotting a nonlinear neural net model that performs well and generalizes



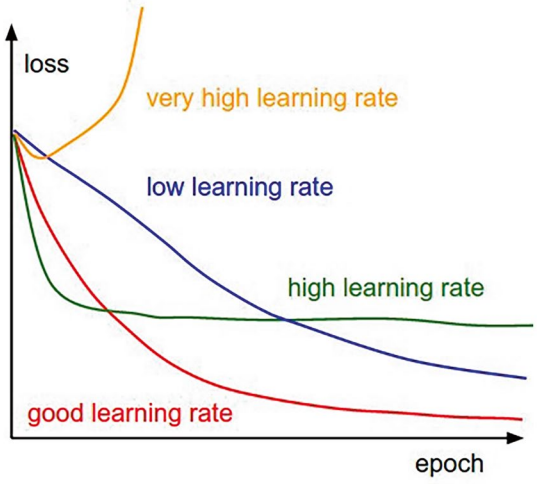


Fig. 31.12 Examining different types of learning rates and their loss rates by epoch

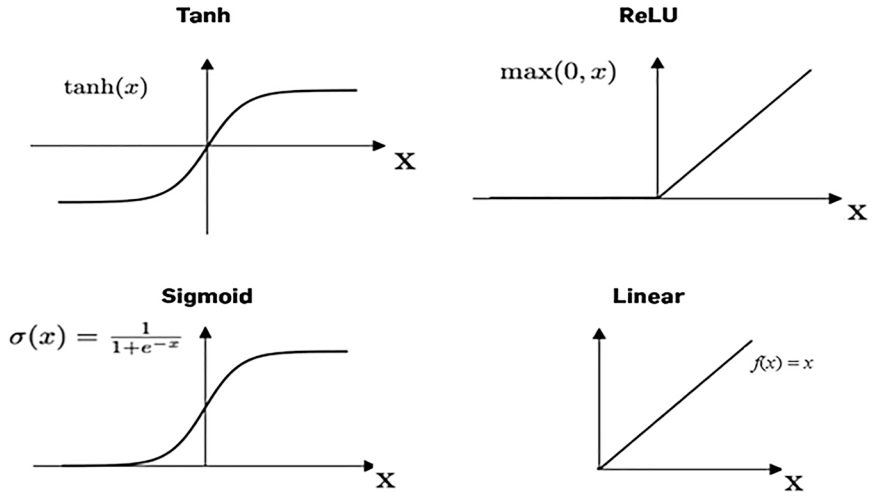


Fig. 31.13 Different types of activation functions

Alongside the activation function, the user selects the optimization method which, in all cases, represents some form of gradient descent. Gradient descent is essentially measuring the slope or change of the variance between the predicted and the observed of two points in the network. When the slope (y/z) becomes 0, then the algorithm has reached a saddle point, either a minimum or a maximum point depending on whether the curve is upward or downward.

In developing deep learning models, hyperparameter tuning is a feature for many of the routines. These routines can have many different options. Hyperparameter tuning, in effect, can perform a grid search on the best option to be used, which can be very computationally expensive. A second approach is to use random search, where only a random number of options or explored. This is computationally less intensive, but it still delivers acceptable results.

It is probable, however, that the most important parameters for the user are the number of hidden layers and the number of nodes per hidden layer. It is here where the user has great latitude in building models that achieve a desired error rate while at the same time being able to generalize on new data.

The power of deep learning and AI has extended to the world of unsupervised learning. The underlying methodology allows the generation of optimum hidden layers based not a label but on another layer. The pattern of data from previous hidden layers is then used to develop subsequent hidden layer data, which are often referred to as autoencoders. These improved hidden layers (or autoencoders), developed from the unsupervised learning of previous layers, can then be better used when required to classify a given image which is now supervised learning.

Reinforcement Learning

Another machine learning and AI technique is reinforcement learning (RL). The basic concept of RL is that you are conducting activities against steady or existing states to move to a desired state and receive a reward. Reinforcement learning is not particularly new, as the mathematics of Markov Models were its earliest applications of assessing steady states and the activities required to achieve a desired state and a reward. With the advent of AI and deep learning, reinforcement learning is now using the concepts of RNN and, specifically, Long-Short-Term Memory (LSTM), where the network places more weights on its more recent results but with the intention of optimizing that reward outcome.

The benefits of RL can be enormous and are being used in applications such as robotics and, more specifically, self-driving cars. In consumer behavior, however, we can observe benefits which are incremental to traditional predictive analytics and which might be those specific business applications. Let's explore this a little more closely. If the ultimate objective or reward in consumer behavior is the maximization of customer profitability, then what are the activities that achieve this goal. In other words, can we use RL to maximize customer profitability? As stated before, most of the work in building any advanced analytics algorithm is going to be concerned with creating the right information or analytical platform. The organization of this platform can be quite complex, especially when there are multiple metrics comprising customer profitability. Think of a bank which generates profitability from a variety of metrics (deposit, risk, lending, investment, etc.). And think of the many desired states that may optimize each metric. At the end of the day, though, the end deliverable is predicting the right activity in order to achieve the desired state of customer profitability maximization. But couldn't a series of predictive models predict the desired activities necessary to achieve the optimal state as opposed to RL and an RNN network which might deliver this optimal state?

The rationale behind this thinking is that, in the realm of predicting consumer behavior, there is much randomness/noise (unexplained variation). For example, the buying behavior of consumers is going to exhibit more randomness or noise than how a car should navigate on a given road. It is this level of randomness in which often the simpler-type techniques (i.e. traditional-type predictive models) will suffice in delivering the same level, or even a better level, of model performance than more advanced deep learning RL algorithms. Why? Because deep learning techniques need a strong signal-to-noise ratio, which is the case for image recognition, NLP, and self-driving algorithms. And, as we have stated before, traditional predictive or machine learning-type models can tend to generalize much better than deep learning models in an environment of strong random noise. This does, however, raise the question of how we might evaluate RL vs. traditional predictive analytics techniques. What might be the framework in comparing a more traditional type of predictive analytics approach to RL? This is particularly relevant for solutions when

comparing both approaches for consumer-behavior predictive analytics-type solutions.

One approach would be to set up a random sample from a customer database and compare it with the balance of customers in that customer database. In both the random sample and the balance, essentially, we are trying to determine the next best action which will ultimately optimize customer profitability. In the marketing literature, much has been written about the next-best action as the ultimate marketing goal. Of course, in setting up the measurement environment, these series of actions would have to be defined in both data sets. Once these actions are determined, however, it is the selection of actions for a given customer that will differ between RL and traditional predictive analytics. In RL, the action is determined based on the output of the deep learning network as it tries to optimize the reward outcome. In traditional predictive analytics, by contrast, separate predictive models would be built for each action. Using some sort of score normalization that looks at expected value, the selected action for that customer would be dictated by the model yielding the highest expected value. Again, the heavy work would be in creating the analytical environment which is similar for both RL and traditional predictive analytics. Once the analytical environment is established, tools are available to quickly create and update the many predictive models that would be required. The use of Auto ML tools to constantly develop new models as well as update them would simulate the automated "black box" environment of deep learning. Yet, one current advantage at least so far in traditional predictive analytics is the ability to explain how the models are working. But besides model explainability and how quickly we can select the desired action, the most important factor in comparing the two approaches is performance. In this type of measurement environment (see Fig. 31.14), we would monitor both approaches as follows:

This would be the report card in really assessing whether or not RL can work in a consumer behavior-type market. Such an approach is not really new as one bank that I worked with in the mid-1990s was trying to determine the benefits of using predictive analytics in their decision-making, as opposed to the status quo which consisted of business rules based on post analysis. They adopted the same approach of setting aside a random sample for predictive

Approach	Average Customer profitability 1 month	Average Customer profitability 2 month	Average Customer profitability 3 month	....	Average Customer profitability 12 months
Traditional Predictive Analytics					
Reinforcement Learning (RL)					
Random Sample					

Fig. 31.14 Measuring impact of reinforcement learning vs. traditional predictive analytics

analytics activities and the balance for the status quo. Given the growth of predictive analytics in the last 25 years, it is unsurprising that predictive analytics emerged as the winner in that time period.

Emerging technologies are all around us, and it is easy to become consumed by the Kool-Aid of not looking like a Luddite. But sound measurement should always be the foundation of new initiatives and activities that we decide to undertake. Many of the experienced data science and analytics practitioners were schooled in this by those company pioneers of data science and analytics. I think it is time to adopt this learned philosophy towards all emerging technologies (including RL), particularly in the area of consumer behavior.

The Impact of AI on Automation of Predictive Analytics Practices

With the evolution of artificial intelligence becoming more mainstream, one may ask about what its impact will be in the automation of those tasks that are relevant to the discipline of predictive analytics. What skills will AI replace within the data scientist's arsenal? In theory with AI, choosing the right mathematical algorithm becomes obsolete as the machine determines the right technique. The machine, through its algorithms, outputs the solution, which can then be immediately applied to a given business problem. In AI, we encounter such concepts as deep learning, which utilize the mathematics of neural nets. As mentioned above, neural nets are not new to predictive analytics and have been used by practitioners for the last twenty years, particularly by marketing departments where analytics is a primary focus. In fact, much of the more recent developments in AI have been about enhancing the developments of neural net algorithms and the literature has provided many examples that have yielded superior results to what was developed 15 years ago. This is particularly relevant in the areas of image and NLP processing. Yet, recognizing that these algorithms are indeed better than in the past, does that translate to improved performance in predicting consumer behavior within a marketing program? The reason for this question, and its specificity towards consumer behavior, is that much of the historical work done in predictive analytics has been in the area of marketing and credit card risk, where we are dealing with consumers as the records of interest. In my experience, much of this behavior is difficult to predict with a high level of accuracy due to the high degree of random error or variation. Under these kind of scenarios, simple solutions trying to generalize overall trends can work quite well with some of the obvious solutions being techniques such as logistic regression or decision trees. But this does not mean that AI should always be discarded when looking at consumer behavior. It simply becomes another tool in the data scientist's arsenal. In fact, software already exists that facilitates the use and exploration of different modeling techniques. There are currently a number of the leading providers in this area to provide this kind of capability.

So where is the biggest impact of AI on data science, and what skills of the data scientist can be done more quickly by the machine? If we look at the four-step approach to building predictive analytics solutions, we can better understand where AI will have the biggest impact. Each of these four stages (identifying the business problem, creating the analytical file, applying the right analytical technique, implementation and measurement of the solution) are critical in achieving an end solution. Can AI completely replace all the skills that are required in each of these four stages, or are there certain stages and tasks where AI will have the most relevance?

In addressing this question, one needs to identify tasks within the process that are repeatable. It is these tasks that AI will replace. Technical competencies in the past, such as programming coding skills in R, SAS, Python, etc., were often the pride and joy of most data scientists. Yet, this once most desirable skill is becoming less important through the deployment of AI tools that can convert a set of business rules into the desired programming code.

The use of programming skills can also be replaced by analytical tools that attempt to replace code with graphical user interfaces (GUIs) and icons that mimic the process and flow of an analytical project from beginning to end.

Figure 31.15 is a schematic of one such process, where raw data is entered at the front of the diagram and then through a series of tasks, we end up with a report which is the final box at the far right of this chart. Note there is no need for programming as the tasks and functions in creating both the analytical file and the required report are represented by drop-down icons.

This ability to facilitate the creation of the analytical file process is, and will continue to be, a key deliverable amongst software vendors moving forward.

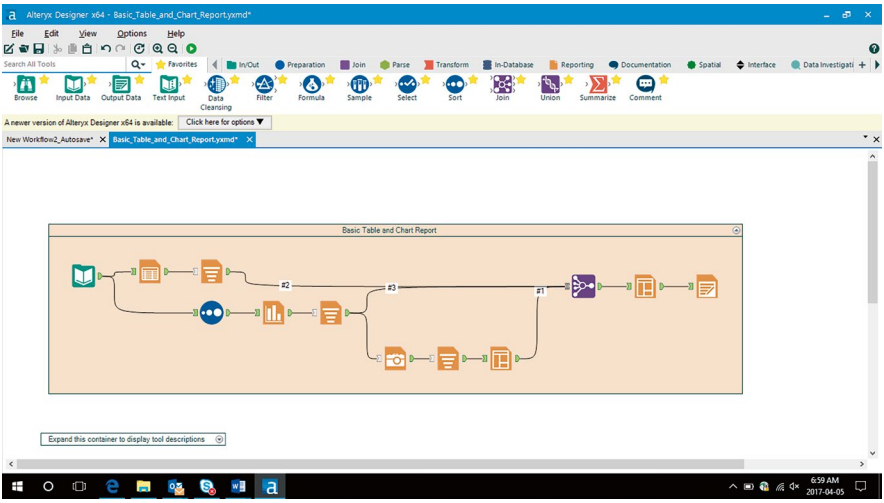


Fig. 31.15 Schematic sample of software and workflow in creation of analytical solution

As a result, the data scientist needs to spend less time in the creation of programming code and more time focussing on aligning the right data to solve the business problem. Gone are the days when the data scientist would simply extract all the data. In a world of Big Data, access to data is no longer a limitation. Instead, the challenge for the data scientist is to be focussed on determining the data that will be relevant and meaningful in solving a given business problem. For example, if I am building a claim risk model, how relevant is the social media commentary related to insurance policies? The relevance in building a predictive model would only be significant if we can match a high portion of these policies back to their social media commentary. Yet, in another example, if we consider insurance fraud, an analysis of social media might be used to detect patterns of communication in terms of geography, where evidence of fraud seems to be most relevant. In both these cases, a stronger level of domain knowledge would help to guide the analyst in what data would be most meaningful. It is this kind of thinking which cannot be automated through any software or AI tool.

The human technical skills of the past in creating the analytical file can now be augmented by software, with the data scientist now focussing on the business problem, the approach to solving the problem and, of course, using the right data. The emphasis for data scientists will be on thinking rather than coding. As a result, more business challenges can now be tackled which, in the past, may not have been addressed due to data limitations or programming resource constraints. Yet, it will still require that the individual have deep knowledge on data, and how it can be used for data science projects. For example, the wide array of procedures and tasks that are used when manipulating data are core areas of knowledge to the data scientist. Bear in mind that data manipulation typically takes up over 80% of the data scientist's time within a given data science project. This will not change, but tools will allow the individual to do this more quickly, thereby allowing more data science projects to be undertaken. Although programming will be less of a need, familiarity with the use of analytics software as well as data will still be a core requirement. Better tools allow the data scientist to focus more on thinking through the problem rather than on programming. In other words, they can focus on what kind of analytical file needs to be created in order to solve the given business problem.

At the academic level, colleges and universities will need to emphasize more of those "softer" skills in training students how to think through a given business exercise. Emphasis on the use of case studies will comprise a large portion of course outlines in any data science discipline. Yet, there will always be a need for those more hardcore technical programmers who can write code and the algorithms to solve the more complex problems, and to ensure the accuracy of their programming solutions. The emphasis of academic institutions in designing data science programs will be directed towards the hybrid. GENAI tools will simply augment the demand for these types of practitioners.

More recent developments in AI, and in particular the use of GENAI tools, are transforming both our social and work lives. The data scientist is no exception here. The use of business prompts will become a common practice for the data scientist.

With academic institutions continuing to evolve their programs alongside the internal training and mentorship programs provided by many organizations, data science as a profession will indeed be bright. Many new developments, techniques, and approaches will emerge, which is a natural outcome of our discipline. The role of the data scientist will continue to evolve as new technologies emerge. Success for the data scientist will be in staying relevant through human thinking that is always external to these technologies.

Key Takeaways

- AI is a game changer in its ability to deliver superior solutions, yet it requires large volumes of data and a strong signal-to-noise ratio.
- AI has been more successful in areas such as image recognition and NLP because of the strong signal-to-noise ratio and the large volumes of data.
- The data science practitioner is like a pilot in adjusting the parameters of the basic neural net architecture to create the right solution.
- AI solutions today encompass deep learning approaches where the neural net architecture encompasses multiple hidden layers, rather than just a single one.
- Developing the best AI solution is about adjusting the parameters such that bias error (overfitting) and the ability of the model to generalize is minimized alongside the minimization of variance error, which speaks to the model's ability to underfit the data.
- Both advancements in GENAI, and the more advanced development of LLM models, are now demonstrating increasing cognition and reasoning and, ultimately, the increased thinking abilities of the machine.
- At this point in time, the human or the data scientist practitioner is more advanced in the cognitive aspects than the machine. This represents an increasing expectation of the data scientist role as a hybrid. The result of this is an increased emphasis on problem solving and creating a data environment to solve that problem and a reduced focus on the technical skills such as programming.

CHAPTER 32

Analytics Today and in the Future

Just how different is Big Data analytics from the original method of data analytics. The internet, and volumes of digital marketing, tend to convey that any topic emanating from this area is seemingly a new one. Seasoned data science practitioners question all the hype and attempt to understand what is so unique and different within the area of "Big Data analytics" when compared with the current traditional approach of data analytics, which has been in existence for more than four decades. The discussion surrounding Big Data Analytics often involves the so-called "five V's" (volume, velocity, variety, value, and veracity) (Fig. 32.1).

Seasoned practitioners have dealt with the volume aspect throughout this period, as well as the V's related to veracity and value. In the era of Big Data, it is the velocity and variety of data that is new and different. Yet the basic 4-step approach and discipline to data science and analytics, which has been discussed throughout this book, has not changed. To remind you, this 4-step approach is as follows:

- Identification of the business problem
- Creation of the analytical file
- Application of data science tools
- Implementation and measurement

So what does the seasoned practitioner, as well as the neophyte need to consider in developing the expertise in Big Data analytics within the context of the 4-step approach. Advances in computer science and search engine technology at Google have resulted in the development of a variety of tools.

From the 1980s, database technology included the development of relational indexes, whereby one would index the match key used to join two tables such as a customer table and transaction table. A good example of such a key

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_32

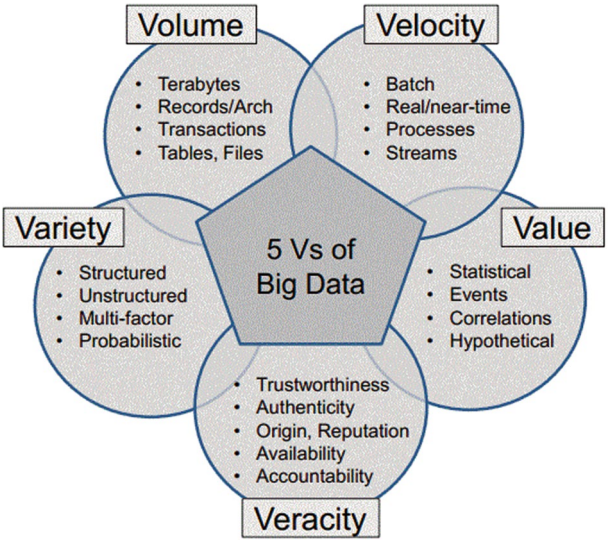


Fig. 32.1 The 5 V's of Big Data

would be the account id. This development increased the data-processing speed, and thereby reduced the time required to join these tables. This became unsatisfactory, however, as customer tables could now comprise several million records while a transaction table might be comprised of many hundreds of millions of rows. This would result in waiting times of 12–18 hours in some cases to simply join these two tables.

The next development was the creation of columnar-based technology, whereby each variable on the database becomes an index. This significantly increased the data-processing capabilities such that the above join between the customer table and the transaction table now took minutes instead of the original lengthy period.

Many people may ask why this wasn't done in the first place. The simple answer is lack of the necessary technology, which was compression. The indexing of a given field greatly increases the storage capacity of that field. Storage became the limiting factor; with compression technology, however, this limitation was greatly lessened, and hence columnar-based technology became viable.

But the technology didn't stop there, as can be seen from the advent of Big Data processing and the use of Hadoop and its Map/Reduce technology. This technology has reoriented data processing from a file sequential-type system to a file distributed processing-type system. This parallel approach to processing data has led to an exponential increase in its speed. Yet this development only addresses one aspect of the five V's: volume. Much of this large volume, in terms of text, video, GPS, and metadata, is unstructured. Seasoned practitioners are very well-versed in dealing with structured data given its longevity of use, but they are also equipped to tackle the unstructured and semi-structured

component of Big Data. The difference today lies in the newer skills which are required in extracting both semi-structured and unstructured data from the internet. Yet, they are typically not equipped in having the requisite programming skills in utilizing Map/Reduce technology and its accelerated data-processing capabilities. Typically, these skills reside in the domain of data engineering.

So what does the data scientist need to understand in this brave new world of Big Data? Let's start with the Map/Reduce technology, which deals with the ability to conduct massive data processing very quickly and does not involve techniques to extract meaning from raw data. There is no real data science at this point and, as stated above, this enhanced data-processing capability is really the resident domain of the computer scientist or data engineer professional. The data scientist does not need to understand all the arcane technical details of Map/Reduce technology but they do need to understand how Map/Reduce technology is integrated in reading and loading data within this new environment. In other words, the data scientist will have to work with tools, or potentially apps, that integrate or provide this technology as part of their ETL (Extract, Transform, Load) toolkit. But a more plausible option is to work with a data engineer that can integrate this processing technology such that the data scientist has the data in a workable format. In working within this kind of environment, here are seven steps that are key to success for the data engineer.

- Create a taxonomy of data classifications.
- Design a proper data architecture.
- Employ data-profiling tools.
- Standardize the data access process
- Develop a searchable data catalog.
- Implement sufficient data protections
- Raise data awareness internally.

One more recent skill of the data scientist or data analyst here is the ability to work with data that is not structured, such as text. Structured data consists of rows or the records of interest while the columns are the actual features or variables. Besides structured and unstructured data, semi-structured data or object-oriented data essentially represent the type of raw data that is captured on the internet. The example below represents the raw data expressed in one tweet (Fig. 32.2).

From the above, you can observe that there are no rows or columns. The variables or actual features are represented as objects. Some examples of these are highlighted in bold. Note that the field denoted as text contains the actual twitter post which is the unstructured component of the tweet. In working within this semi-structured environment, the use of Python is the better and more commonly used programming language in extracting this raw type data and converting it to the structured format of rows and columns. But the tools

StatusJSONImpl{createdAt=Wed Apr 16 08:48:20 PDT 2014, id=456459080618749952, text='RT @GoT_Tyrion: Looks like Catelyn put @JonSnowBastrd on the far side of the window #GameOfThrones http://t.co/uIN7u6IOgk', source='web', isTruncated=false, inReplyToStatusId=-1, inReplyToUserId=-1, isFavorited=false, isRetweeted=false, favoriteCount=0, inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=319, isPossiblySensitive=false, isoLanguageCode='en', lang='en', contributorsIDs=[], retweetedStatus=StatusJSONImpl{createdAt=Fri Mar 14 09:27:09 PDT 2014, id=444510052452298752, text='Looks like Catelyn put @JonSnowBastrd on the far side of the window #GameOfThrones

Fig. 32.2 Example of object-oriented Twitter record

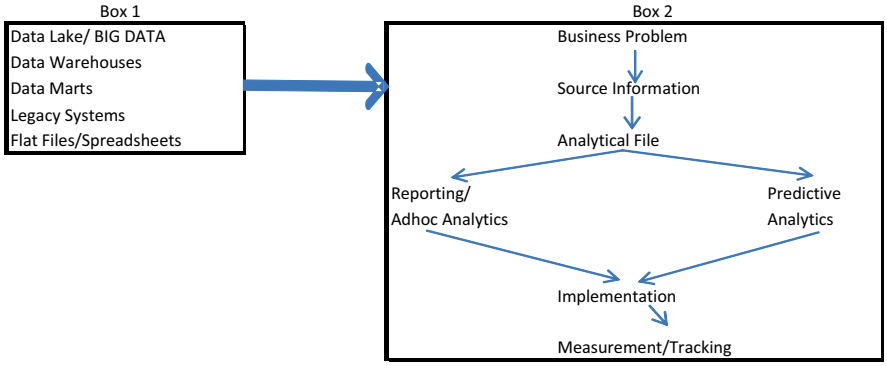


Fig. 32.3 Integration of Big Data processing and analytics

and processes of data analytics and data science can only operate with structured data.

Figure 32.3 depicts the integration of Big Data processing and analytics.

Some 90–95% of the data scientist's work is represented in Box 2. Yet, organizations and industries have devoted hundreds of millions of dollars to Box 1 as data lakes and processing techniques related to Big Data all occur in Box 1. From a data science perspective, this, of course, is important as access to technologies and new information simply offer the potential of erecting better solutions. But organizations have unknowingly devoted disproportionate larger resources to Box 1 when, in fact Box 2, which is the mainstream approach to data science, represents a huge knowledge gap for most organizations. Before going into more detail about Box 2, another aspect of Big Data is the notion of first-party data vs. third-party data.

In today's digital and Big Data environment, no one disputes the value of first-party data. Customer tombstone (i.e. demographic information) information, purchase behavior information, and touchpoint (clickstream) behavior represent the more traditional types of first-party data that is collected by most organizations around a given customer. Examples of third-party data include mobile and social media data alongside website browsing behavior. With all this data, we can build a very robust analytical file. Of course, this all assumes

that match rates are high, with the resultant analytical file being representative of the customer group we may be trying to promote. The secondary concern is one of privacy and whether the ability to append social media, web traffic, and mobile behavior information violates the privacy laws, which were discussed in Chap. 9. Third-party data can be significant, but the current regulatory environment is erecting roadblocks around the use of this type of data. In other words, the ability of Facebook and Google to analyze my behavior based on ads posted on their platforms by companies might be severely restricted.

How much literature and content is focused around the concepts illustrated in Box 2. The data science discipline does itself a disservice by not elaborating in further detail the actual mechanics of each topic which is illustrated in Box 2. Box 2 is the essence of real data science. Accordingly, this book has discussed at length the importance of creating an analytical file and the process of generating the appropriate input features for a given data science solution. This is not accomplished by hiring computer engineers, who know how to code in Java or Python so that the data lake can be converted into a structured file or semi-structured format that can still be used for analysis. The concept of building the analytical file remains. The ability to manipulate this data and create meaningful input variables is still the domain of the data scientist. Java programs, or "Big Data" computer engineering technical skills, represent one component, albeit an important one, but certainly not the overriding consideration when considering its value within the discipline as represented by Box 2. This is not to diminish the importance of certain "technical skills" being critical to extract key elements within this "data lake" environment (see arrow from Box 1 to Box 2). Yet, it is just one component and clearly a minor one within the context of building the overall solution as depicted in Box 2.

Using the analytical file to conduct the right analysis or develop the appropriate models is the next step; this is followed by implementation and measurement/tracking, which again have been discussed in this book. All these concepts mentioned in the above paragraph are in Box 2, and have been discussed at great length in this book.

It is irrelevant whether I use Python or SAS or other tools that fundamentally allow one to enact the process outlined in Box 2. Tools are merely enablers, albeit important ones; they are only one small part of the overall puzzle. But we are looking here at the process and approach in conducting the right analysis to solve the right business problem. It is equally important to consider the following: have we thought about how to implement and measure these solutions? Too often, I hear about supposed data science solutions, but which lack the rigor and discipline of a key measurement process that proves business value incrementality. I once asked a question at a conference to a senior executive of a major bank that was using operations research methodology as their marketing optimization approach. My question was whether they validated the impact of this approach. His response was a puzzled look of bewilderment with the attitude that "measurement in this case was irrelevant as we had to trust the 'math'".

As data scientists, we have a long way to go in both educating ourselves, and also, more importantly, the "business community", in what we do particularly in a world being increasingly driven by generative AI (GENAI) technology. Focusing our efforts on Big Data processing and extraction technologies does our discipline a disservice. Instead, we should focus to Box 2 and spend more time in expanding our knowledge base in this space. Case studies and business examples should include scenarios involving simple math as well as some involving more complex math. Much like the legal profession in common law, where much of the knowledge is captured through a case-oriented approach, data science needs to expand its knowledge base beyond just the technical and mathematical arcane tasks and instead focus on how this knowledge is applied within many practical business settings.

The process and functions of data analytics and data science are no different from what seasoned practitioners have been doing for many decades. The picture below (Fig. 32.4) provides a template describing the process of Big Data analytics. Concepts such as sampling, data aggregation, reporting, feature engineering, training and testing, and the use of advanced statistics, were all common topical areas in the discipline of data analytics prior to the internet (Fig. 32.4).

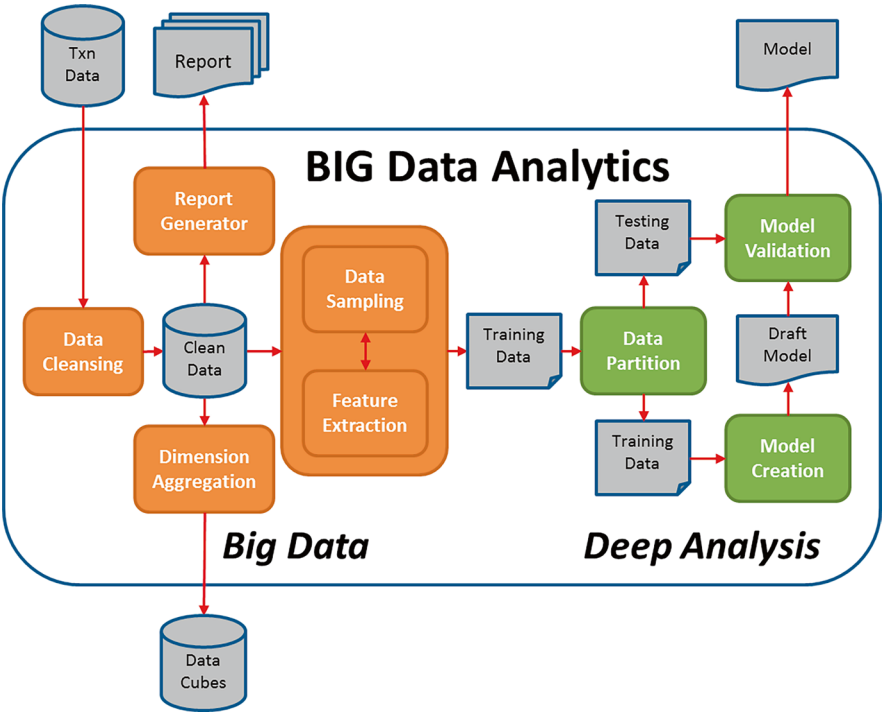


Fig. 32.4 Big Data analytics: Is it different from small data analytics?

Although Big Data has indeed mitigated the significance of sampling in analytics, sampling is still a viable technique. There are many different styles of sampling, with the most common being random sampling, where a random number is assigned to each record that essentially allows you to extract whatever proportion that you desire in each data set. In other literature, random sampling is often referred to as bootstrapping. Nth sampling is a very similar technique to random sampling, in which one extracts every nth record from the original data set. However, one cannot assume that nth sampling is always random, as there may, in fact, be some inherent structure to the order.

A third approach, stratified sampling, is a technique used in building predictive models to predict a probability, such as response to a campaign. In this case, the target variable comprises binary values|:1's (person responded) and 0's (person did not respond). In some cases, the 1's are very few with many 0's and, in this case, the data scientist will grab all the 1's but only a proportion of the 0's, hence the attachment of the name: stratified sampling. See the schematic example of stratified sampling in Fig. 32.5.

Similarly, boosting could be considered an offshoot of bootstrapping or random sampling, but this method differs in the sense that certain records may get reweighted in the sample. This can be done when building models as more emphasis or weight is placed on records that are misclassified (Fig. 32.6).

There is still that missing link in being able to create an analytical file (both for advanced and non-advanced exercises) which is often forgotten or not discussed by the Big Data vendors. It is this analytical file that provides all the potential data inputs which can be used as potential model variables or key cluster segment variables, or variables within a massive pivot table or report. The missing link, or analytical file, is the key to creating meaningful business solutions. One typical example might be six million customer records complemented by several hundred million records transactions. In such a scenario,

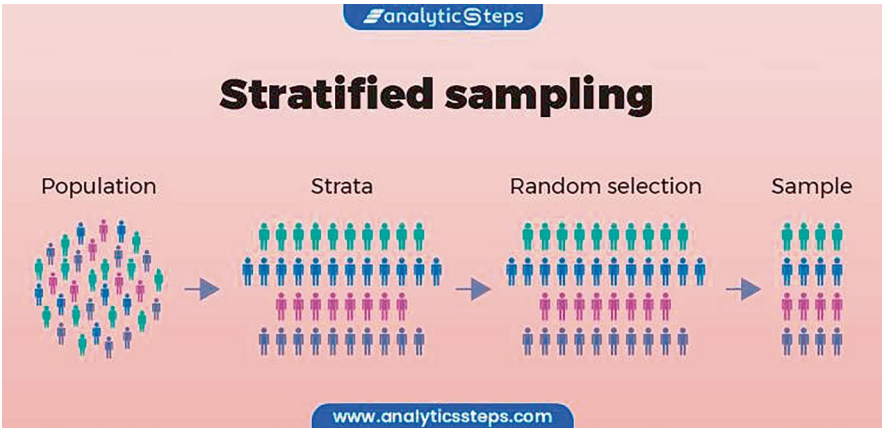


Fig. 32.5 Schematic of stratified sampling

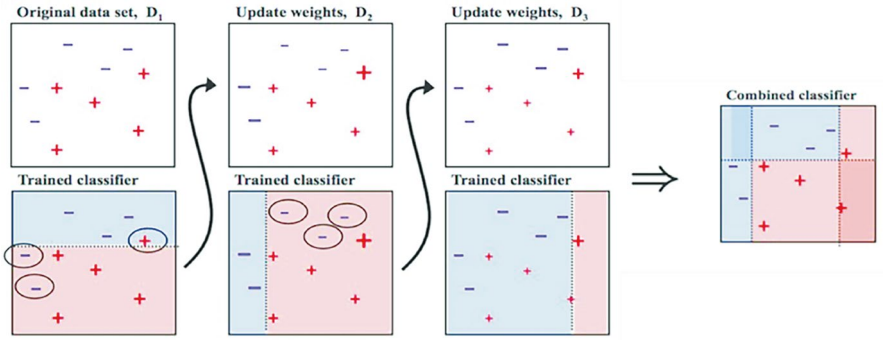


Fig. 32.6 Reclassification of records using boosting techniques

how does the data scientist create the analytical file? Hard work is the quick answer. Initially, this entails extensive data audit processes and routines which have been discussed in this book and are used by the data scientist to fully understand each data file, and also each data element within a given file. This process allows the data scientist, for lack of a better word, to become "intimate" with the data. It is this knowledge that lays out in detail how the data should be worked in creating the analytical file. This would reveal the following:

- What data elements or variables should be included in the analytical file?
- How should the data files be merged?
- What source variables are useful?
- What new or derived variables do we need to create
  - Binary
  - Aggregation/summarization
  - Change-related?
- What additional external sources (not collected by the company) can we append to the above information?

The above process, which I have outlined, may seem simplistic. Yet, it is not unusual for hundreds of new variables to be created in this process.

One of the real challenges in developing predictive models in the marketing area is in the area of acquisition. The reason is the lack of data. In thinking through this dilemma, actual information is quite limited and confined, in most cases, to name and address information. Certainly, the use of Stats Can data in appending geo-demographic data to prospect data through the postal code has historically been successive. But are there other data sources that could be used to help improve these models? One creative approach is to use existing customer data as input variables in any acquisition model. The key here is to have sufficient customer data as the customer data at the individual level is not being used at an individual level, but is essentially aggregated or

summarized to a geographic level. For example, variables such as average tenure or average spend are based on customer behavior, and many others could be created at a geographic level. More data allows the creation of these variables moving towards a more granular level, such as a six-digit postal code or even at the less granular forward area (the first three digits of postal code). Our experience has shown that the use of internal customer data at the geographic level, in conjunction with Stats Can data, has improved the overall lift of acquisition models. As stated above, however, this improvement, and its associated level of geographic granularity, will depend on the volume of customer data that is available to us.

Social media data offers an ocean of data that can be used to create meaningful analytical variables. In social media, where much of the data is either unstructured or semi-structured, the practitioner adopts the same approach: Given a block of unstructured or semi-structured data, what transformations need to take place to transform raw social media information into meaningful variables. Within social media, however, the newer challenges for the practitioner reside in how to extract these key fields within a varying format. Programming tools and apps continue to improve and facilitate this extraction process. The real difference, or new challenge, is to ultimately convert this unstructured data to structured data where the analytics can be conducted. Once this extraction process is complete with a structured table as the output: summarization/counting/aggregating/categorization routines can then be used to create meaningful data science variables which is exactly the same process used against raw transaction data. The real difference between social media data and traditional data sources prior to the advent of the internet involves the mining of actual text content. This is a whole discipline unto itself involving a number of processes that allow the user to ultimately create meaningful variables from the corpus of text.

Like text mining, sentiment analysis can be conducted against the "data" in order to identify the "emotional" value of the text. Here the practitioner's ultimate objective might be to simply quantify this emotion on a scale of 1 to 5, with 1 being most negative and 5 being most positive. In all cases, though, the analyst is "working" the social media data to derive the right information as part of the analytical file. Knowledge of the end analytical goal will dictate the data extraction process i.e. the requisite key words in extracting the relevant social media commentary.

But let's think more about the context of this data. As data scientists, social media is certainly voluminous and messy but, as has been often stated before, do we still really need this information? Once again, the definition of the business problem or challenge (first phase of the data science process) will determine the utility or usefulness of social media. The phrase of "one size fits all" does not apply to social media. The volume and variety of data requires us to be very focused on what information we actually need in order to solve the business problem. It is the nature of the business problem that will dictate how the analytical file will be created. For example, are we simply trying to create

reports to conduct some kind of trending analysis, or are we trying to build a model? Recognizing how we want to create the analytical file, marketers need to understand the context of the information. For instance, how representative is the information concerning the target audience? Are people who comment on brands/products representative of the population at large? This is a fundamental question to ask when conducting analytics. In science, this rigor has always been applied to any analysis, as the core data is always vetted in terms of a bias factor. Prior to social media, we observed this rigor to some extent since data science practitioners were always aware of the so-called "responder bias" with market research surveys. If models were being built, they were always applied against the same biased group. Insights and learning on biased research that could be used for communication purposes were utilized cautiously. This is why data science has focused on the most unbiased form of information, which is what customers have done through their exhibited behavior with the organization. It is observed behavior that marketers want to better understand. Crossing over into the digital divide, the so-called "responder bias" may still come into play, as these individuals may represent people who like to be heard but who do not represent the viewpoints of the silent majority.

Yet, for marketers, the real "holy grail" is to tie observed behavior to actual purchase. Whether this is done digitally in a Big Data setting, or in the traditional database systems, it does not matter. The analysis of actual observed behaviors to predict future outcomes always yields the best solutions. Data scientists will, of course, look at everything, but will have a more laser-like focus on data that focusses on actual observed individual-level behaviors which can be tied to an individual outcome when constructing the analytical file. It is this approach that more readily lends itself to outcomes that can easily be measured from a return on investment (ROI) standpoint, and to the ultimate driver of corporate behavior, profitability.

Let's take a look at some real practical uses of data analytics in our Big Data world. Perhaps the most important and more relevant perspective of Big Data today is the ability to act on the information in real time. However, much of this does not involve complex analytics. For example, I am in a particular transit station where I receive an alert on my phone regarding the nearest Tim Horton's. I then go to the nearest Staples store, where I receive another alert on my phone indicating that a certain product is being offered at a discount. I then get into my car and then receive an alert to slow down based on my current driving speed.

In all the above three cases, Big Data is being consumed in a simple analytical fashion and used to drive action. But no real advanced analysis or even trend analysis is being conducted. Instead, an action is being delivered based on one data point. In the transit case, it is simply the destination of the nearest Tim Horton's to that transit station. In the Staples case, it is simply the fact that I am in the store which has the product discount. Finally, the speeding alert in my car is predicated based on my current drive speed. As one can imagine, there are endless activities that can be suggested to the consumer based on the

data point or activity currently being exhibited by the individual. Use of the current available information rather than analytics is being used to drive the decision.

In the transit example, instead of just telling me the location of the nearest Tim Horton's, the use of higher-end analytics could supplement this message by referring to products that I am most likely to purchase at that time of the day. Similarly in the Staples example, in addition to offering me a discounted offer, the store can recognize me as a high-value customer and offer me more discounted offers. In the driving example, the use of this additional behavioral information can be used to offer more flexible pricing arrangements regarding my premium. The incentive to adjust my driving behavior more quickly in order to reduce my premium would be a primary motivator in improving my driving behavior, at least in the short term.

Consider another hypothetical example. Suppose me to be a passionate Toronto Raptors basketball fan. Since my entire relationship in terms of all my purchase behavior (tickets and non-tickets) is being captured at the individual level, all kinds of product offers and discounts can be offered to me based on what I have done in the past, but also, more importantly, on what I am expected to do in the future. Talk about the "WOW" factor, as the Raptors are not only engaging me with a brand that I love; they are also engaging with me personally on specific products and services that are more meaningful to me.

It is this ability to utilize a deeper level of analytics, complemented with real point-in-time-type data that provides a much more enhanced understanding of the customer. The development of predictive analytics solutions from this "deeper type" analytics is the real "Holy Grail" of analytics in a Big Data world.

Model Explainability

The other issue with using advanced analytics such as AI in a Big Data environment is explainability, which is, in effect, a very significant barrier. Most organizations want to gain a better understanding of what is driving a given consumer behavior, such as customer response or credit risk or attrition risk. Data science practitioners have historically been able to explain the key components within a consumer behavior model and its relationship to the desired consumer behavior. For example, as discussed in Chap. 12, traditional machine learning algorithms like multiple regression or logistic regression can look at the statistical measures, such as the partial R^2 of a given model input variable, and compare it to the total R^2 of the entire model in order to obtain the explainability of that variable. Listed below is a table of a predictive model which we observed in Chap. 12, Fig. 12.1. This table as discussed is attempting to maximize the likelihood of a bank customer responding to a credit card product. This table depicts the contribution of each model variable based on the partial R^2 relative to the total R^2. It also indicates the impact (negative or positive) on the desired modeled behavior of response (Fig. 32.7).

Model Variable	Impact on Response	Contribution to Overall Equation
Behaviour Score	Positive	35%
Average Score	Positive	25%
Have an RRSP Product	Negative	15%
# of Fin. Inst. Products	Positive	10%
Avg. % of Credit Limit Used	Positive	10%
Live in Prairie Provinces	Negative	5%

Fig. 32.7 Final model variable report

The use of a deep learning model poses explainability challenges due to the complex nature of the solution, with all its hidden layers and nodes. Much of the current research in using AI techniques, however, is in this area of explainability, and there are some interesting approaches which can at least convey the importance or contribution of a given variable, but not necessarily the sign or impact. Some of this research considers how much overall scores change through the elimination of a single given variable. The change in score becomes the key metric in assessing a given variable's importance. Another option, particularly in image recognition, explores the importance of a variable or input within a certain area of the photo or image. Essentially, it measures the intensity of the pixels of a given section within that image.

ML (Machine Language) and NLP (Natural Language Processing) Empowerment

The emergence of Big Data has led to considerable growth in the use of text for text analytics which has been discussed in previous chapters and Natural Language Processing (NLP) where text is utilized as data to generate textual or voice responses. We observe its common uses today through chatbots, with Siri (Microsoft) and Alexa (Google) being perhaps the most common.

The advancements in data science and, in particular, NLP now empower organizations to leverage what people are saying to the point where the machine or computer now responds to the customer using GENAI-type techniques, with CHATGPT being the most common one. The computer, through NLP processing and the advent of Big Data and AI, has accelerated the level of solutions where the machine now learns on historical text and utilizes that knowledge to generate a response which is specific to the question. We are even becoming familiar with this in domestic settings, through the use of "smart devices."

These types of interactions (which are AI-driven in terms of the machine response) is now used across virtually all industry sectors. The difference today

for many organizations is the level of sophistication that is being used to analyze this text.

Although there will always be benefits in interacting with a live human, the machine and GENAI tools are rapidly improving. The future challenge will be how to become more proactive, using the machine as a tool or enabler with the human as the expert guide or pilot. Yet, the real downside or risk is that the reverse occurs, with us becoming so reliant on the machine that we become the tools and the machines become our guides or pilots.

The Hybrid Role

At an even deeper level, what might be the hybrid role today here for the machine and the human? Could we marry technologies such as NLP and ML with the agents' skills and knowledge to provide an even better service to the consumer? For example, analytics could be used to identify what customers are likely to do such as "Are they likely to go bad, how profitable are they, and are they likely to leave us?" ML-based predictive models, which represent common business practices, are the tools underpinning access to this type of information.

The competitive advantage for organizations today lies in utilizing both types of tools, allowing the use of ML to predict consumer behavior and its resultant prescriptive outcomes, complementing this with what the consumer has said, and then determining the appropriate form of communication. It is in such interaction that the agent supersedes the machine through this type of complementation. Let us look at one example.

Suppose I call up the customer service center at RBC Royal Bank of Canada. The agent receives my identity-related information and knows that I am a high-value customer but one who is likely to defect based on ML models. My previous conversations with the bank now provide the agent with information, through NLP and text science techniques, on how to communicate with the customer more effectively. The agent may determine, through knowledge of previous dialogs, that this high-value, likely to defect customer may have a specific interest in wealth products. Using this information, the agent might emphasize the unique features and benefits of those available from Royal Bank.

In another example, a customer is deemed, through the use of ML models, to be more likely to be at a higher credit risk. Using the above tools of text science and NLP, previous communication dialogs indicate interest in more lending-type products. Once again, armed with this knowledge, the agent might suggest a plan to pay down debt with the goal of having access to those types of products.

Data scientists and ML practitioners with any degree of business experience have always advocated the use of both the human and the machine in any of their analytical efforts. Today, it is this collaborative hybrid approach, which marries technology with people, which is the key to success for all organizations. And this collaboration is now an imperative in a world that is increasingly consuming the benefits of GENAI technology.

The Real Accuracy of Consumer-Behavior Predictive Models

Perhaps the biggest myth underlying predictive analytics is its ability to accurately predict outcomes, particularly in the area of consumer behavior. The mathematical perspective in assessing outcomes related to model accuracy is typically related to diagnostics like R^2 for techniques, such as regression or other similar diagnostics. But let's see what this really means from a practical perspective.

Suppose we build a response rate model, in which the average response rate before the use of any predictive model is 2%. If this model is 100% accurate and is being deployed against a universe of 100,000 names, we would expect that the top model-scored 2000 names would all be responders ($100{,}000 \times 0.02$), yielding a perfect R^2 value of 1. The experienced practitioner will never observe this, however; if this indeed were the case, then grave concerns would be considered regarding the massive overstatement of the model. In fact, a more likely scenario would be that the model achieves a 6% response rate in the top 10% of model-scored names, where 30% of all observed responders have been identified in this top 10%, and there was a model R^2 value in the range of 0.05. Certainly, 30% is better than 10%, which would be the case if the solution were completely random. Yet, the mathematical purist would still be correct in saying, "30% of all responders is good but what about the other 70% of responders?" Indeed, what about this other 70%? Can we not do anything to capture more of this 70% group, thereby improving the R^2 beyond 0.05? The answer to this question, in one word, is "NOISE."

Predictive analytics, and the use of statistics, is about explaining trends and attempting to identify patterns. The implicit understanding from a mathematical perspective is that these tools explain the variation in the data. Models that yield better results in theory explain the total variation of the data. A perfect model R^2 of 1 implies that the total variation of the data can be explained. This is never the case. There is always some unexplained variation that represents our "NOISE."

In the consumer marketing arena, and also, to a lesser extent, in consumer credit risk, the seasoned practitioner perspective is that very high degrees of noise or unexplained behavior are the norm rather than the exception. Our challenge as practitioners is to accept this reality, yet to use these tools and our knowledge within this environment to yield significant dollar benefits. As any mathematician or statistician will tell you, our tools are about explaining trends and patterns, and, in effect, reducing the noise.

In the world of business, and indeed marketing, this ability to truly explain away this noise is severely limited. It is not uncommon to have models or tools that explain only 5% of all this noise, with the result being that the model is performing well above average when compared with other models. But why do we accept these so-called "dismal" statistics, such as 70% of responders not being accounted for and 95% of the noise not being explained? The answer to

this is that significant incremental dollar benefits are accrued in explaining only a small portion of this "NOISE" or variation. The real issue when considering "NOISE," however, is identifying what is true noise rather than noise that can be effectively explained away through a predictive model. Unfortunately, this is not something that a mathematics textbook can really explain in terms of its practical impact on business. It is only by applying models in practice that one observes this huge disparity between that noise which can be explained and that which cannot. Here the use of validation techniques on a holdout data set is extremely important in attempting to isolate noise that can be truly explained.

Concepts such as splitting one's analytical file into test, training, and validation data sets allow the analyst to observe how the model would have performed against a completely new set of data. Of course, there are several techniques that can be pursued here, with the most common being the use of proportions of the analytical file to create the data sets. Cross-validation is another technique in which an analytical file is split into several parts. Let's say we divide this analytical file into five parts. The first part is used to develop the model, and the second part is used for validation in the first iteration. In the next iteration, the second part is used for testing and the third part is used for validation. This process continues until we have used all five parts; where there are five models we then take the average of the models. This is to be regarded, in a sense, as an ensemble modeling technique.

Another approach in assessing models is to use data sets from different time periods in which the model was originally built. This provides the analyst with more reinforcement concerning the model's performance and the model's ability to generalize on future unseen data.

In the real world, predictive analytics are often used to predict outcomes where the observed "yes" cases are indeed much fewer than the "no" cases. For instance, response and defection models all operate under scenarios where the "yes" outcomes typically occur in less than five cases out of a hundred. The extreme proportion of "no" outcomes, by its very nature, creates more noise. A good example of this is that the model's ability to accurately predict outcomes increases as the rate increases to 50%. In fact, a good exercise is to try to build models where the rate is close to 50%: one will observe that diagnostics explaining the power of the model (R^2) increase when compared with models where the rate is small (5% or less). We attempted this exercise ourselves and built a predictive model using the same data fields and the same target variable, which, in this case, was response rate. Figure 32.8 depicts our results, which support our hypothesis that increasing response rates translate to increased model power or performance, as demonstrated by R^2.

Having talked about the ability to increase overall model performance through increased rates (that is, response, defection and so on), we observe that even in our experiment the model performance, or R^2, peaked at 7.24% with a response rate of 50%. But we can easily ask the questions, "What about the other 92.7% that the model is not able to explain? Is this true random noise, or can we further explain some of this noise?" The key in attempting to

Table 1	
response rate	R^2
0.60%	0.34%
2.76%	1.13%
5%	1.55%
50%	7%

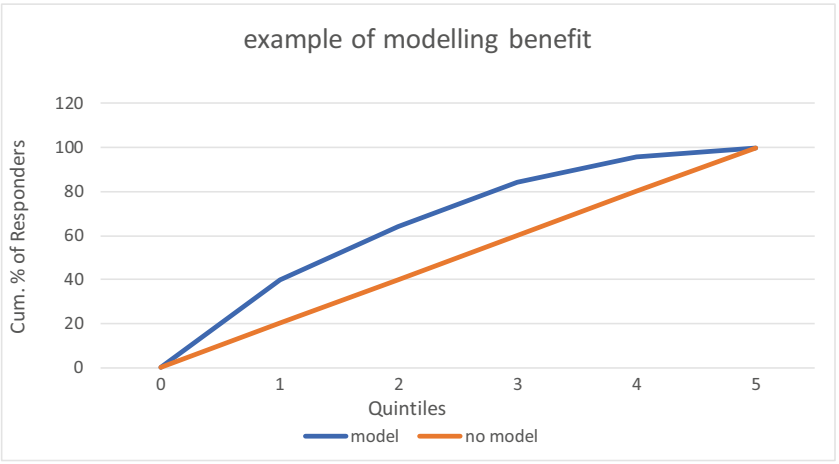
Fig. 32.8 Increasing accuracy of R^2 as response rate increases

explain away more noise is to identify variables that exhibit nonlinear relationships with the target model variable, yet which make sense regarding the behavior we are trying to predict and the business where these solutions are to be deployed. The use of pure mathematics without an understanding of the business could result in models that yield nonlinear-type variables that are simply attempting to explain away true random noise. As discussed above, the notion of having several validation samples—one that is derived from the same analytical file as the development sample and another that is derived from a different time period—can certainly help to mitigate model overstatement caused by excessive nonlinear-type variables. The use of these validation data sets will filter out those nonlinear relationship-type variables that are indeed true noise. Yet, even in undergoing this type of rigor in enhancing our model performance, it is unlikely that we can reasonably expect models to explain more than 10% cent of the target behavior. However, if high levels of noise represent the typical modeling environment, and modeling inaccuracy for marketers is the norm, then why do we consider these techniques? **PROFIT.** The best way to answer this is through a case study.

In the case study presented here, we are looking at a response model that optimized the likelihood of purchasing a premium-type product. The response model itself comprised five variables with an R^2 of 0.05, indicating that 95% of the variance is unexplained, or represents pure noise. The figure below (ROC curve), as discussed in the literature, represents our first view of model performance (Fig. 32.9).

The table above represents the cumulative % of responders at each decile. Plotting these numbers creates the curves for the model scenerio vs. the non-modeling scenario. This generates the AUC curve, which is the area between the model curve and the no model curve (Fig. 32.9).

The businesspeople would still reply, "So what?," as their real interest would be the actual dollar benefit of building a predictive analytics tool. Figure 32.10 below exhibits the benefits of targeting through predictive analytics against not using the model. This represents the cost savings achieved by promoting fewer names to achieve a desired level of responders. In this case study, the promotion costs were US$1.00 per effort. As seen, if the desired quantity to be promoted is 200,000 names, the incremental promotion costs required to achieve 3200 responders without a predictive analytics tool is $120,000. This means that it is the $120,000 that is the real measure of success. Despite a low R^2, who cares if, in one campaign, the business can achieve $120,000 in benefits?



	0	1	2	3	4	5
model	0	40	64	84	96	100
no model	0	20	40	60	80	100

Fig. 32.9 Example of modeling benefits using cum. 0% of responders and AUC curve

Table 3: Example: Benefit/opportunity of predictive analytics			
Business Scenario	Number of Responders	Number of Names required to achieve 3200 responders	Promotion costs
With Predictive analytics targetting the best 200000 names	3200	200000@1.6% response rate	$200,000.00
Without Predictive Analytics	3200	320000@1% response rate	$320,000.00
Dollar Benefits			$120,000.00

Fig. 32.10 Demonstration of $ benefits of modeling

However, multiply this impact over multiple campaigns and one can easily see the huge impact of predictive analytics in environments where the tools are mathematically not that accurate. So, as an old business mentor of mine used to say, "Let's ground the statistics to business rather than the business to the statistics." It is this grounded philosophy among most present-day practitioners that has led to the embracing of these techniques by the business community.

Data Analytics in a Post-COVID World

In this new post-COVID-19 environment, it is not unrealistic to assume that the way consumers behave, and think, will be transformed significantly. Of course, this has ramifications when conducting analytics exercises. Virtually all data analytics exercises deal with historical and longitudinal data. The development of models, segmentation systems, and/or reports all use historical data in their solutions. Given that much of the power of predictive analytics/machine learning solutions arises from longitudinal or historical data, this all begs the question of how we deal with data, and specifically consumer behavior data, prior to, during, and after the COVID-19 crisis.

As I pondered this issue, I was reminded of the last crisis that seemed to galvanize our collective consciousness and made us reflect on what is truly important in our lives: 9/11. During that period, this newfound awareness did have an impact on consumer behavior. My organization at that time was asked multiple times about the impact on our analytics exercises, especially the more advanced analytics exercises such as the many predictive models that we had built. In other words, the question was posed: would model performance be significantly compromised because of these changes?

Our perspective in evaluating model performance was to use our classic gains chart/decile table approach, as outlined in Chap. 13. Using this approach, rank ordering of observed target behavior against modeled names becomes our measure of success. Figure 32.11 gives a graphical representation of a gains chart of a response model to target existing customers in the purchase of an insurance-related product.

In this above example, we observe the same response model being applied during normal times, and then immediately after 9/11. In this example, the

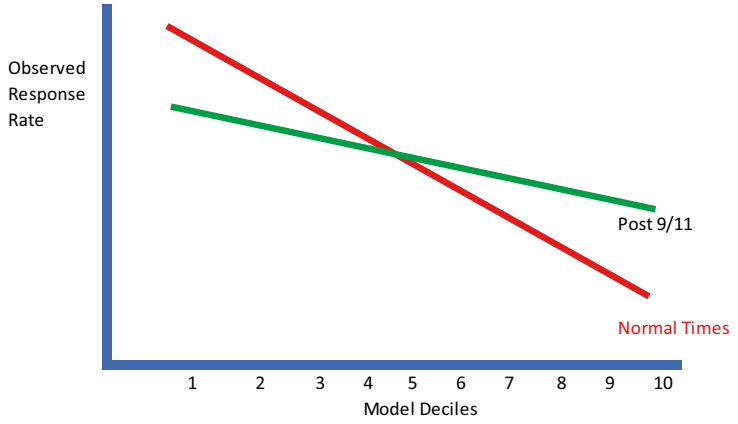


Fig. 32.11 Example of model performance changing under extreme conditions

model has eroded as the slope of the line has decreased, indicating the model's reduced rank-ordering capability.

Figure 32.12 shows another insurance response model example of a situation where again there has been a change in performance before and after 9/11.

Yet, in this scenario, we observe that overall response rate performance has deteriorated in the post-9/11 period, but the model itself continues to perform as the slope of the line remains the same, indicating no decrease in the model's rank-ordering capability. This scenario is highly unlikely, however, as model erosion is probably the most likely scenario after some major crisis which impacts consumer behavior.

These above examples serve to highlight the fact that analytics solutions can change quite dramatically, and for the worse, both from an overall perspective and from a modeling perspective. The erosion of model performance during and after a crisis is somewhat intuitive, as consumer behavior is not continuous throughout the population. Certain characteristics about consumer behavior, such as their age, where they live, income, etc., may all be impacted differently by an extraordinary situation. As data scientists who are always trying to mathematically assign weights to these variables based on patterns in the data, this becomes a flawed option as we know these data patterns have now undergone significant change.

So, what are the options for analytics practitioners in such an unstable environment? For many experienced practitioners, the philosophy of simplicity has always been at the core of what we do. Capitalizing on this notion of simplicity, one of the first concepts of basic analytics is the use of indexing, in which a specific consumer behavior or demographic of an individual is compared to that same consumer behavior or demographic of a group. A classic example of

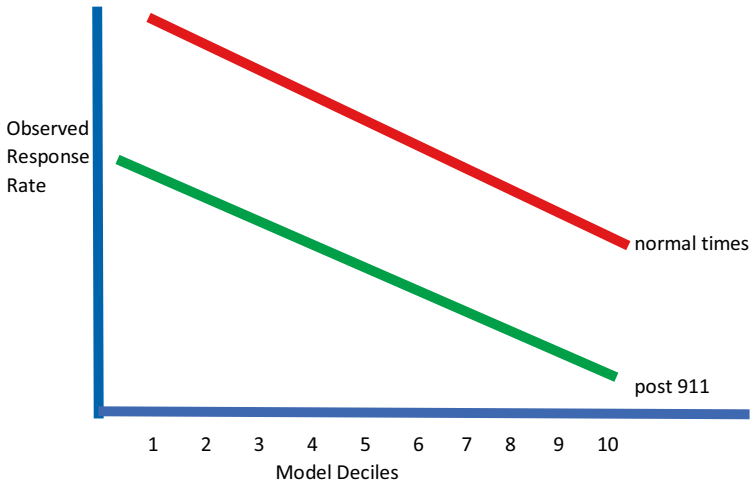


Fig. 32.12 Example of model performance not changing under extreme conditions

	Recency of last activity(R)	Frequency of Activities(F)	Amount Purchased per Activity(M)
Customer A	2 months	30	40
Group	4 months	20	50
Index	2	1.5	0.8
RFM Index for Customer A	=1.46(2 + 1.5 +.8)/3		

Fig. 32.13 Calculating RFM index

this is the use of RFM, where each consumer is indexed on recency of last activity, frequency of activities, and the average amount purchased for a given activity. This was discussed in Chap. 14. Listed below is an example of how RFM might be calculated for one customer (Fig. 32.13).

In this simple example above, we are assuming that each behavior (R, F, M) is equally weighted in calculating the overall index of 1.46. The use of an index adopts a comparative approach in analyzing data. Although the RFM ranks and indexes can change dramatically during an extraordinary event like COVID-19, the real objective in this exercise is to determine the relative difference in RFM scores and ranks during this period of time. This can help in better understanding migration behavior. In other words, how have customers changed their actual behavior during this tumultuous period of time?

We have used this type of indexing or comparative approach multiple times and not always when there has been an extraordinary global event. For example, many organizations will undergo significant changes in their corporate strategy, which will have a dramatic impact on consumer behavior and the accompanying data environment. In these situations, we have developed indexes based on customer value and change in customer value or behavior change, which was discussed in Chap. 18. In many of our consulting engagements, this was our recommended segmentation approach due to a data environment that would be dramatically shifting over time.

Of course, in a more stable data environment, this indexed approach is suboptimal as the advanced machine learning and deep learning technologies can really leverage the abundance of data that is now available at our fingertips. As we enter this new post-COVID phase, however, analytics practitioners need to be very cognizant of unstable data environments and changing consumer behaviors. As I have conveyed to fellow practitioners and students for many years, data science is about identifying patterns in the data. But if these patterns are undergoing significant changes due to external events, then a more simple and pragmatic approach such as indexing can prove a very useful targeting tool in a very dynamic data environment.

Analytics in the Mobile World

The key to conducting effective data science is, of course, data. And that has historically been the perennial challenge for many companies. The advent of loyalty programs and e-commerce platforms have certainly helped in this area through the creation of the necessary platforms to gather and collect this rich information at a customer level. However, many organizations do not have loyalty programs or e-commerce platforms as a data collection mechanism. This situation is commonplace in the retail, restaurant and energy sectors. But mobile marketing analytics has become the game changer here and presents itself as an option for data-starved companies. It is not actual customer purchase data but location data or customer visit behavior, where again we assume that higher and better visit behavior does translate to increased purchases. However, let's examine this visit behavior more closely.

Each phone has a unique MACID, which essentially acts as a pseudo-customer ID. The phone has a built-in GPS, which can essentially track our location. Through our phone, we can track visit-type behavior. Based on the strength of these signals, any business with a given location can determine when the customer entered and exited the establishment, thereby creating information based on length of stay, or "dwell time."

The actual data collection would involve the data scientist, creating a geofence or radius around a given location. The data analyst, along with the business stakeholder, would determine the appropriate radius level. For example, should it be 100 yards, 500 yards or even a kilometer? Of course, this will all depend on the business and its mechanics when drawing in customers. Obviously, the radius is only the first component. Individuals can enter a radius and then quickly leave it, as they are simply walking by the establishment. So, how do we identify a genuine visit? We might decide that a visit begins when a phone entering the radius remains stationary within that radius for a proscribed period of time, where the duration is determined by the business.

But how would the data scientist actually gather this data? By working with certain data providers such as Uber Media, Bell media, etc., the data scientist conveys the necessary parameters, such as the radius or geofence around the given location, the length of stationary time within the given radius, and the time period used to extract all the data which, in many cases, comprises millions of records.

In addition, this information can all be analyzed at the most granular level, which is, of course, the individual phone. Marketers, of course, could use this data to send personalized messages. Yet with stricter data governance protocols now in place, marketers should and would need to obtain consent in conducting any one-to-one marketing of this type. But analytics of this individual mobile phone data is different and can be analyzed in a non-intrusive manner. Through this rich granular-level data, RFM measures, such as recency of last visit, frequency of visits, and average duration of visits, can be determined along with time of day, weekend vs. weekday, or even type of season. With this

data, we can determine the trends of our best visitors compared with our worst visitors in terms of RFM, allowing us to see if there are emerging trends or insights based on the time of the day, the day of the week, or the season. For instance, in a restaurant, we might find that our best visitors now seem to be coming more often in the morning. Maybe this trend is due to our changing breakfast menu. From this information, we might decide to more actively promote this menu during the afternoon and evening. Alternatively, maybe it is the issue of speed that drives this increased engagement, given people are in a rush to get to work. A speedier delivery format for lunch might thus increase engagement for those lower-ranked RFM visitors.

Other valuable information from this mobile data might be demographic in nature. But how can we gather this information when the only data received by the phone is really location data and the timing associated with this data. This is where data scientists have again expressed their creativity. We know that the location data contains latitude and longitude (lat/long) data, which can be mapped to postal codes. With postal code data, we can then map this information to Stats Can data in order to arrive at the demographics, such as age, gender, income, ethnicity, education, etc., that pertain to the current postal code existing on the phone. But what we really want to know is where the phone or mobile user lives. If we look at the lat/long data of the phone between 12 a.m. and 5 a.m. and observe that the phone is stationary, we can assume that postal code to be the residence of the mobile user. As with all aspects of data science, this is not perfect as people could be working night shifts. For the most part, however, it is a reasonable assumption, and we can now append this rich demographic information to the phone. For example, what do the best RFM visitors look like based on where they live? Maybe they comprise younger people with higher levels of education but average incomes. Armed with this information, we might decide to promote the health aspects of our menu. The restaurant could initiate campaigns that promote the creation of social groups that advocate the nutritional value of our products. Again, however, the key here is to act on this insight in order to further engage our visitors.

Another game-changing application is the retail store. These can use beacon technology, which facilitates the collection of information as the customer and their phone move around the store. Analysts can then derive where a customer goes in a store on every occasion they pay a visit. Let us consider how the retailer could improve each customer's experience upon entering a store. If historical navigation data indicate the customer frequently visits the cosmetics aisle, the retailer could dispatch messages, special offers, and coupons relating to the products sold there. Similarly, someone who previously shopped for workout gear could be offered a coupon for running shoes. Beacon technology can transform the shopping experience, making it more relevant and engaging for each individual customer. The assessment of trends regarding movement of customers throughout the store could enable it to better align its display of goods with regards to customer movement.

The issue of data privacy, as previously discussed, must also be paramount here. Are we being respectful of that basic individual right? This will, of course, depend on how the data is used. If it is being used for analysis to detect trends and patterns for analysis, then this is not an issue. But if we are using it to market to individuals, then the issue of privacy must be addressed clearly to all consumers.

Future Thoughts: The Big Data Discussion and the Key Roles in Analytics

Big Data. Hype and more hype. But what is the reality here? Is big data the real deal, as many experts would have us believe? For seasoned data science practitioners, big data has always been the norm rather than the exception. Building predictive models at American Express in the late 1980s exposed practitioners to a million cardholder records with hundreds of millions of transaction records. The challenge, then, was to translate this raw source information into a meaningful analytical file that could be used as inputs for developing predictive models. At that time, marketing databases and data warehouses were new to practitioners. The use of crude tools, such as legacy mainframe systems, required a certain level of technical knowledge concerning the use of JCL (job control language) and the storage of data. Accessing data and data output on tape or disk required technical knowledge to optimize the use of data science in this big data environment. For instance, in creating analytical files, it was not unusual to have files with hundreds of variables. However, the analytical file was created through intense programming by practitioners. As with any programming routine, these programs had to be tested rigorously in order to ensure that the programming objectives of the analytical file are being achieved. To do this effectively, the testing and quick feedback of the results represents the desired outcome. However, with a million records and hundreds of millions of transaction records, the data would be stored on tape, and the actual programming routine would be submitted as a batch job, with the results being relayed the next day, not in seconds. To conduct effective testing of the programming routines used in the creation of the analytical file, practitioners would extract a small random portion of the raw data and store it on disk. By doing so, practitioners could test their programs on the sample data and determine whether programming changes had to be made. With this approach, the analytical file could be created in a more timely manner, even with the crude tools available back then. This approach to using random data would also have been employed in stage 3 of the data science process; namely, the stage involving the use of statistics in the development of the model.

As the American Express case study shows, the ingenuity with which practitioners dealt with information in a Big Data environment was key to building effective data science tools in a timely manner. As the PC environment evolved and marketing databases and data warehouses were introduced, the improved

technical infrastructure of big data significantly advanced the development of data science as a discipline. Yet, despite technical improvements, data volumes have continued to explode due to the growing significance of CRM and the development of the internet. The world of social media has further accelerated this explosive data growth. Are larger data volumes the real challenge for practitioners? The answer would be a resounding "No." What are the challenges for practitioners in this big data world?

First, let's explore three of the 5 V's: volume, variety, and velocity. As discussed, the volume component is not a new phenomenon for practitioners. Yet, the other two Vs do represent challenges. What do I mean here? Variety represents information and data that is arriving in different formats. This scenario is best exemplified in the unstructured and semi-structured data that is the norm in social media. Blogs, videos, posts, text messages, and emails are all examples of unstructured data. Most practitioners are very comfortable in the big data volume environment if the data is structured. However, what happens when the data is unstructured? New tools, technologies, and skills are required in order to best optimize this information for data science purposes.

The old approaches of extracting information in structured rows and columns no longer apply. Tools that can extract pods of information rather than rows and columns and programming techniques that extract the critical information now allow analysts to work effectively in this new paradigm. Platforms such as Hadoop Map/Reduce allow organizations to process data more effectively through what is called distributed file processing. Then, a number of tools, such as NoSQL, Python, Java, and many others, provide the programming languages that allow meaningful information to be extracted. The critical task for data scientists is to determine the best option to extract data and convert it to a variable or variables that will be meaningful for a given business objective.

The emphasis on big data and big data analytics has focused the data science discussion on the importance of the data itself, rather than on the mathematics. Once again, this scenario goes back to the basic 4-step process of data science; that is, the process begins with determining the data science challenge or problem and then, in the next step, constructing the appropriate analytical file—and doing this in the big data environment. Most experienced practitioners welcome the discussion surrounding Big Data because it has increased the awareness of data science and, more importantly, of the process of creating the right data environment. New terms, such as data science, now testify to the importance of data in the data science process. Data science means that practitioners undertake a rigorous and methodological approach in dealing with data. From a practitioner's standpoint, is any of this new? Data as a discipline has always been the core foundation of developing any data science solution.

The discussions around big data have also led to supposed new roles in the world of Big Data. It is amusing to see long-standing roles or functions being repositioned in the frenetic world of Big Data analytics. For long-time practitioners, a certain cynicism arises from role repositioning. The regression

analysts of the early 1980s became the modeling analysts of the late 1980s, which then turned into the more commonly accepted term of data miners, and have now been more exaltingly defined as data scientists. Even disciplines such as direct marketing analytics are now encompassed under the general term Big Data analytics.

Let's now take a closer look at the roles of data scientists and business users who are involved in any data analytics project. Back in the 1980s, these roles existed in many leading direct marketing companies. Direct marketing companies had key individuals or modeling analysts, called data scientists by today's definition, who were building direct mail response models. Yet, such models could only be used with the involvement of marketers, commonly known today as business users. This required a marriage of the two disciplines: data scientists had to have some knowledge of marketing; and the business user had to have some knowledge of modeling. Yet, this so-called crossover knowledge would be irrelevant without deep domain knowledge being paramount for each role.

To better understand these two roles, it is useful to see what this means in direct marketing, because that is where the history of building data science solutions began. In the early days of analytics, the discipline was almost exclusively used in a tactical manner. The overall marketing strategy was determined by the key leaders in the marketing area. Specific tactics, such as marketing campaigns, were then designed and created to execute this strategy. In order to ensure that these campaigns were most profitable, analytics were conducted in order to optimize efficiencies, and these were measured in such performance indicators as reduced cost per order or reduced cost per customer saved. In this very tactical exercise, these two roles (data scientist and business user) are distinct, yet complementary in achieving their objectives. The marketers (business users) have usually already developed the campaign or initiative and its accompanying budget. The optimization of budgets was the primary driver of analytics in the early days. That is, the analytics might provide insights into developing basic business rules that would reduce costs. For example, in acquiring new customers, the identification of key lists or particular Stats Can demographic segments was essential. At this level, marketers were simply trying to understand the key characteristics differentiating new customers from the general population. Over time, the cost reduction from this learning would erode, which means that a more advanced analytical approach beyond basic business rules was needed. Marketers wanted to develop a model, and so they engaged with data scientists. The business users or marketers would explain the objective of the campaign, and that would usually define the required modeling tool. Listening carefully to the marketers' campaign objectives, the modeling analysts would then think about which data to assess. For example, what data sources do they use in building an acquisition or retention model? To develop a model, data scientists also must consider whether they can create the necessary information environment that will deliver meaningful inputs for a model, and whether they have the data to create the objective function or target variable. More importantly, they must assess whether they have the software that

allows them to create the information or the analytical environment and to then leverage the software by being able to run a variety of statistical routines.

In today's Big Data context, analytics is a much more collaborative process, and marketing and business strategies are much more data-driven. It is common for organizations to commence their foray into analytics with a data discovery, as discussed in Chap. 29. Data discovery is a process to define a data strategy that addresses an organization's business needs. The collaborative approach to data discovery helps determine the necessary tactics for the execution of the strategy.

As the importance of Big Data has grown, so have the roles of data scientists and business users. With social media and text data being readily accessible, the ability to understand context, and those business insights that might be meaningful, will remain the purview of business users. Meanwhile, the ability to manipulate and derive meaningful information which is necessary for these insights is the responsibility of data scientists. These roles will continue to grow in importance as big data and analytics become increasingly entrenched in a company's corporate culture. This will result in more collaboration and a more symbiotic relationship between business users and data scientists. Without this close collaboration between these professionals, companies will be at a disadvantage in this new age of information and big data.

Companies can gather data in ever-increasing amounts, but how can they make sense of it all? The data is often disconnected, and, as raw or source information, this data is often meaningless unless it is converted into meaningful information.

Even structured raw data must be transformed into meaningful information. For example, postal code information by itself is useless for data science purposes. However, grouping postal codes into categories, such as regions, suddenly transforms this data into more meaningful variables for analysis. The classic example of raw structured data converted into meaningful information is transaction data. Here, data scientists create and derive different kinds of transaction behavior variables. Raw transaction data can be summarized in countless permutations from basically three fields:

- Amount
- Date
- Type

In addition to nearly infinite summarization options, data scientists or data scientists can also capture change. For example, with the date field, analysts can identify how behavior is changing over time and link this change to the type of transaction.

The ability to link data is essential. For example, information on online behavior, purchase/transaction information, campaign history, and demographic behavior all needs to be linked to gain a more holistic view of that individual or customer. Here, technical skills are required for determining the

relationship between files and tables and the way to link the tables. Often, analysts must construct the fields to link these tables, a task that requires technical expertise.

Transforming raw data into meaningful variables is complex and hard work. What are the key ingredients to achieve success? Going back again to our 4-step data science process, the first step is to identify the business problem. Then analysts must work their magic with the data. The essence of this magic derives from two areas:

- Domain knowledge of the business and of what is required for business to be successful
- Mapping the current data/information environment into meaningful variables derived on the basis of the above-mentioned problem or challenge

The emergence of more technology and software tools, and particularly AI, has expanded the data science field to individuals with minimal knowledge of programming. Yet the discipline of data science relies on a foundation of knowledge that is programming- and software-independent. For example, the rigour and discipline which is utilized in transforming raw data into meaningful data inputs or an analytical file represents a deep base of knowledge that comprises a significant component of any data science program. As mentioned earlier in the book, according to leading practitioners, some 85%–90% of a given project's work deals with data. Another critical piece of knowledge is deep knowledge in statistics. Knowledge of calculus and linear algebra is also a growing requirement due to the increased practical use of AI and deep learning.

From a practical perspective, however, it is about the ability to understand the mathematical and/or statistical output and, more importantly, what it means to the business. This requires extensive knowledge in how to properly measure the performance of solutions and in understanding such technical concepts as the overfitting and underfitting of solutions, which are often referred to as bias error and variance error.

The underlying theme that resonates from this preceding discussion, and from the whole of this book, is the ability to gain data science knowledge that will not be replaced by either existing or future software. This implies that training should place less emphasis on technical skills such as programming, and more on the practical knowledge that is required to be a successful practitioner. Tools and technology will continue to improve, resulting in reduced demand for specialists and their technical prowess. Instead, the need will be for that data science generalist who has the required depth of knowledge when it comes to data and mathematics/statistics, but who also has the breadth of knowledge to apply it to an infinite array of business problems. The demand for these generalists will focus on their ability to think through a business problem. These generalists will become like "chess masters" as they align the right tools, the right data, and the right mathematics in solving the right business problem.

Key Takeaways

- Big Data is nothing new. The same 4-step process discussed throughout this book still applies
- The focus still remains on what to do with the data once it is extracted, involving processes such as data hygiene and variable creation.
- Big Data has introduced different structures of data (unstructured and semi-structured) which can now be analyzed. In this new data world, real-time streaming data is now relevant for analytics exercises.
- Model explainability is a more significant issue in AI solutions.
- The issue of unexplained variance or "noise" will continue to be an issue, particularly with consumer behavior predictive models.
- It is still the case, however, that models with low levels of explained variance can still yield tremendous business benefits
- Simpler solutions suffice in a more dynamic data environment
- Mobile technology has expanded our analytics capabilities to include location and visitor type analytics
- With programming becoming a less critical resource, given the increase in GENAI tools, data science continues to evolve more towards the role of the generalist

CHAPTER 33

The Do's and Don'ts of Data Science

Tip 1: Identify the Right Problem

Going back to the 4-step approach to building data science solutions which has been discussed in detail throughout this book, the reader will recognize that this tip is the really the first step of our 4 step approach. As any business student will tell you, this issue is prevalent throughout all business disciplines. Yet, companies continue to face difficulties regarding this step. These difficulties arise because organizations encounter many business problems in all facets of their business. As discussed previously, however, not all these problems require a data science solution; more importantly, how does one prioritize these problems. Effective identification of business problems requires the expertise and knowledge of the data science discipline, complemented with the domain knowledge of the specific organization and its industry. We have discussed the need to gain more domain knowledge of the business as a core requirement in being able to identify the right problem. In this information-gathering process, the first step is to identify those key stakeholders who will be most impacted by the project. Interviews and meetings will be conducted with each of these stakeholders. In many cases, though, this process is simplified as there may be only one key stakeholder who, in fact, has been designated to represents the interests of the company. In any case, a picture of the real needs and issues from the stakeholder's perspective begins to emerge. At the same time, the stakeholder is also sharing their domain knowledge with the data scientist, thereby providing insights to the data scientist on what solutions might be appropriate.

But this process does not end here. Reports and analyses are also sought out because the knowledge and expertise of the data scientist in analyzing these reports may provide new insights and findings that may otherwise have gone unnoticed by the company.

R. Boire, *Data Science for Managers*,
https://doi.org/10.1007/978-3-032-19705-4_33

Once this information-gathering process is complete, the issues and needs become much more focused towards identifying problems that can be solved by data science techniques. For example, if I have a campaign which yields a response rate of 2% without any data science, and which has a required breakeven response rate of 20%, there are other issues beyond data science that need to be resolved here. However, if the breakeven response rate was now 5%, the potential gap is more reasonable, thereby allowing us to consider potential data science solutions. Deeper exploration of these issues through a more discovery-type approach can identify these and other issues which will better highlight the real business problem.

Once we have effectively identified these issues and needs, some basic-level analysis should occur that allows organizations to prioritize these issues and needs based on the opportunity cost of not undertaking this data science work. This can then allow the organization to more closely align its resources towards those areas with the largest payback. Realizing that data science is about maximizing profitability through increased cost effectiveness, one very simple approach is to simply multiply the cost of the record by the volume of records. For example, if I have three initiatives (A, B, C):

- A: A retention campaign to a potential audience of 100,000 customers with a promotion of $1.00 per customer
- B: An acquisition campaign to 1,000,000 customers at $0.25 per prospect
- C: A upsell campaign to 50,000 customers at $10.00 per customer

Using the simple formula (Cost per record X Volume), the priority would be:

C:	50,000 × $10 = $500,000
B:	1,000,000 × $.25 = $250,000
A:	1,000,000 × $1.00 = $100,000

This is just one approach, but it is a pragmatic one that yields some basic insight in prioritizing data science initiatives.

Tip 2: Getting Stakeholders Onside

Like any discipline which is project-based, one of the keys to success is collaboration amongst the key stakeholders. How do we achieve a solution that is embraced by all the key stakeholders within the organization? First, we need to understand that the path to achieving this is not necessarily straightforward as roadblocks on this path require the individual to think in a more collective rather than an individual vein. What do I mean by this?

In building solutions that are more collaborative rather than individualistic, our first step is to identify the key stakeholders within the project. Typically, these stakeholders will certainly comprise individuals from departments such as marketing and IT. In more complex-type projects, this can include finance and

operations. A more complex type of project might include solutions which involve the creation of customer-level profitability.

Projects that are deemed to be more strategic rather than tactical will also include more stakeholders across a larger variety of departments. The key in determining these key stakeholders, though, is to identify those individuals who will be most impacted by the solutions.

In a strategic type of project, such as the creation of customer-level profitability, many individuals are impacted by the solution. Besides the ability of marketing to develop and execute campaigns that are more focused on return on investment (ROI), finance can also be more precise on the impact of any business decision based on customer profitability, and how it impacts the overall P&L. In terms of executing this strategy, employees who are on the front line in terms of serving customers, are equipped with better information about that specific customer which, ultimately, translates to a more appropriate customer-level experience.

For more tactical-type projects such as building a predictive model, this effort might consist of the data scientist dealing with a marketing manager and a person from IT. In fact, the level of involvement from IT might be very peripheral as this department is only used to explain that information from the database which the data scientist does not understand. Much more interaction would occur between the marketer and the data scientist as the marketer needs to understand the solution. This does not imply that the marketer needs to necessarily understand the arcane technical mathematical details behind the model; at the very least, however, they should understand the information or variables which comprise the model. Using the explanation of the solution being a "black box" because it involves advanced statistics is not an acceptable reply. A more common understanding of the solution between marketers and data scientists helps to foster a culture where the model becomes a common component of the business process. As we are all aware, most models will be successful, but some will indeed fail. Adopting a "black box" approach prevents a more collaborative effort in attempting to find out why the model did not work. In this so-called "black box" approach, the data science practitioner can explain mathematically what is going on, but they may not be able to communicate what are the real drivers and key variables of the model and, more importantly, their importance within the model. This is fine if the model continues to perform. However, if the model is unsuccessful, both the practitioner and the marketer have extreme difficulty in explaining the causes for failure. The causes for model failure are best determined by "opening up the hood", so that both the marketer and the data science practitioner know the key model variables, their relationship with the desired modeled behavior, and their overall contribution within the model. With the adoption of a "black box" approach towards modeling, this type of analytical forensics cannot occur. This can ultimately lead to increased resistance towards the use of future data science solutions.

When creating either a tactical or a strategic solution, the key is communication amongst all the stakeholders as this leads to increased understanding and, ultimately, more likely buy-in towards a given solution. But the ability to achieve a certain level of understanding that creates engagement and buy-in is no easy task. This can require much effort and work on the data scientist, but the rewards can be great. Through this approach, the data scientist demonstrates the ability to collaborate with others to achieve a given solution. Corporate executives and business books identify this trait as one of the keys to being a successful leader in today's business environment.

Tip 3: Creating the Quick Win

As with any new process or change, there will always be resistance, or simply a desire to carry on as before. The simple introduction of data science as the nirvana of increased ROI does not necessarily imply that it will become a standard business discipline within the organization. Employee engagement and executive-level engagement still adhere to the phrase of "Show Me the Money" from the Tom Cruise movie *Jerry Maguire*. For many within any business organization, it is the immediate direct $ impact of any change which will ultimately alter behavior. Data science is no different. But how do we demonstrate this impact?

Establishing quick wins is the key. Producing a solution in a relatively short time frame with minimal resources can yield results which do indeed address the issue of "Show Me the Money". This can be as simple as finding an initiative where no prior targeting has been carried out. Using a simple index approach such as RFM (Recency Frequency Monetary), we then apply the RFM index to an upcoming initiative. We can then select our targeted group, that which has the highest indexes, and then compare them to a control group which has been randomly selected. Listed below is a list of customers (500,000) that have been selected based on their RFM index.

$ Benefits of RFM				
	# of Promoted Customers	Resp. Rate	# of Responders	Total Cost
With RFM	500000	3%	15000	$750,000
Without RFM	750000	2%	15000	$1,125,000
Difference	250000			$375,000

Using a promotion cost of $1.50 per customer, the RFM index yields a saving of $375,000. This saving is achieved since marketing can promote fewer customers with the RFM index to achieve the same number of responders (15,000) that are achieved without the RFM index.

These kinds of results can also be demonstrated without waiting for a future initiative. If we can identify previous initiatives that have yielded results without the use of any targeting, we can then apply the RFM approach to this initiative. We can then determine what would have happened if we had used this approach in selecting the best names. This approach is often referred to as "back testing" and it is both quick and very cost-efficient one, with the only cost being the practitioner's time in conducting this type of analysis. Armed with these types

of results, the practitioner and marketer can more easily engage the use of data science in future initiatives, and ultimately begin embedding it as a standard business discipline.

Tip 4: Understanding the Data

This tip represents the core of all data science exercises, and it applies to all media. The benefits of any data science exercise are only going to be as good as its inputs or data. Without a clear understanding of what comprises these data inputs, misinformed insights, as well as faulty results from any subsequent analysis, will yield decisions that are not only sub-optimal but, in some cases, detrimental to the organization. Organizations need to have a discipline that manages this process, with the result being the ability to conduct sound analytical exercises due to a solid understanding of the data.

In Chap. 6 "Creating the Analytical File", we discussed the data audit process as our methodology in becoming more "intimate" with the data. Variables and fields are filtered out of the process due to issues related to missing vales or cardinality. The process also conveys how the tables or files should be linked together, i.e. one to one, one to many or many to many. Insights from the data audit also reveal to the practitioner what derived variables might be useful in a future analysis. The ultimate end objective, however, is an analytical file where all the fields are well understood by the practitioner and can then be used for the next stage of the process: the analytics stage.

Tip 5: Judicious Use of Statistics

The blind use of statistics without understanding what it means at a practical level lead to the likely outcome of unactionable learning. This implies not only that the data scientist understands results but also that this understanding extends to the business user involved in the project. For example, correlation reports should be able to communicate the relative importance of a given behavior against a desired behavior such as response. In a way, these types of reports should be able to describe what the desired person looks like. The correlation report (Fig. 33.1), based on response as the objective function, would tell the following story about what a responder might look like:

Fig. 33.1 Correlation table vs. target variable of response

Variable	Correlation Coefficient	Confidence Interval
Tenure	0.2	99%
Total Spend	0.15	99%
# of products	0.13	98%
Age	0.12	97%
Income	-0.1	95%
Credit Score	0.002	12%

- Has been a customer for a long time
- Tends to spend more
- Has bought a large number of products
- Is older
- Has lower income
- Credit score has no impact on response behavior

In building models with these above variables, both the statistical significance of these variables and interactions between these variables are important in determining the final model variables, as are the parameter estimates or weights attached to each variable. For example, the report (Fig. 33.2) indicates that tenure is, by far, the strongest model variable (accounting for 60% of the model's contribution) and obviously much stronger than is indicated by the correlation results. So what is going on? Interaction effects between all the variables, commonly referred to as multicollinearity, causes the resulting equation to comprise three variables, with tenure accounting for a significant portion of the model's power.

As seen above, the model results convey a slightly different story to those shown by the correlation results. This is acceptable, however, as the results will be used differently for marketing purposes. If I want to target specific names, I am going to use the model to obtain the right names. But if I want a more complete description of what a responder looks like, correlation results would provide the more relevant type of information.

As mentioned in the previous tip on "Understanding the Data", some fields or variables might contain numerous outcomes that are character format. This would suggest that we create binary "yes/no" variables for each outcome. Yet, some of these variables would have too many "yes/no" variables, with very few records having a "yes" value. There needs to be a way to group these outcomes into more meaningful categories. CHAID, a statistical routine which is often used to build predictive models, can be used in this context to create these categories. This was discussed in Chap. 18.

Product sequencing and affinity analysis also employ statistics. Similar tools, such as correlation, are used but, in these kinds of applications, different types of correlation reports are produced to provide the necessary solutions. These types of analytical exercises are very data-intensive and result in multiple iterations of correlation-type reports. The iterations are produced to present

Fig. 33.2 Final model variable report

Model Variable	Impact on Response	% contribution to Model
Tenure	Positive	60%
Total Spend	Positive	22%
Income	Negative	18%

different perspectives of the data. For example, we may want to look at product sequencing and affinity behavior by value segment in order to see if there are changes in these types of behaviors across these segments.

Advanced segmentation uses techniques such as clustering, with the result being to produce distinct customer segments whereby each one exhibits the same behaviors and characteristics. Any analyst building cluster solutions will tell you that there is both an art and a science in arriving at the right number of clusters, as well as in describing the characteristics of each cluster.

Frequency distribution reports (Fig. 33.3) are the next set of reports which reveal how the values of a variable are distributed amongst the records in our file. The reports would indicate, for instance, that tenure was not reported prior to 1998 as well as 43% of customers having no start date, and that product B seems to be the most prevalent product amongst this group of records.

Figure 33.4 represents a type of report which is continuously generated every time updated data is used for an ongoing analysis. The report below is a standard KBM (Key Business Measure) report that explores significant changes with regards to some of the metrics and to identify potential problems in the business. This information can then identify or prioritize those areas that are in need of a more exhaustive analytics-type exercise.

Tip 6: Combining Art and Science

All experienced analytics type practitioners will agree that there is a definite blend of art and science in creating optimum solutions. Within each exercise or project, there are specific moments where the analyst looks beyond the science in arriving at approaches in building better solutions. For the sake of brevity, I will look at two examples.

The first example is the adoption of the adage "Fish Where the fish Are". To the practitioner, this might imply that a good place to commence any analytics is to explore those areas which have a high level of customer penetration. More specifically, within the area of acquisition, exploring the notion of existing customer penetration as a means of targeting new customers certainly represents one sound analytical approach. Future acquisition efforts would then focus on the selection of prospects that reside in high customer penetration areas. Of

Tenure	# of Customers	% of Customers
1998	9800	14%
1999	10000	14%
2000	12000	17%
2001	8000	11%
Missing	30000	43%
Total	69800	100%

Type of Product/Services Purchased	# of Customers	% of Customers
Product A	35000	29.66%
Product B	40000	33.90%
Product C	25000	21.19%
Product D	15000	12.71%
Other	3000	2.54%
Total	118000	100.00%

Fig. 33.3 Frequency distribution reports

Fig. 33.4 KBM measures over time

	Period 1	Period 2	Period 3
Record Count			
Average Purchase Amount			
Average Age			
Average Tenure			
Etc.			

	Income	% 3+ Household	% Landed Immig.
Average Postal Code	$40,000	52%	5%
M5A 1J2	$50,000	60%	10%
Index	1.25	1.15	2

Fig. 33.5 Example of art and science in creating business decisions without customer data

course, depending on the level of sophistication and the data environment, other approaches would also be used in conjunction with the "Fish Where the Fish Are" approach. This type of approach has been used quite extensively in the not-for-profit sector.

A second example is what do when no customer data exists. Even the use of the "Fish Where the fish Are" approach would not even apply here. So, is there anything we can do in this situation? Closer examination might reveal that market research information has been conducted which indicates that the purchasers of this company's products have higher income, have a large number of persons in their household, and are immigrants.

The experienced analyst would realize that, although customer data is unavailable, there is Stats Can information which relates to these characteristics, albeit at a postal area level. Using this knowledge, the analyst can create a composite index of these three characteristics which are listed below (Fig. 33.5).

This composite index or score for a prospect residing in M5A1J2 would be $(.3 3 \times 1.25) + (.33 \times 1.15) + (.33 \times 2) = 1.45$. This kind of indexing approach could then be used to score all postal codes, thereby providing some means of targeting prospects within the targeted postal codes for acquisition-type programs that essentially leverages the market research learning. Although there is something of a leap in faith in the sense that we are assuming that market research behavior can be inferred on database behavior, it is better than doing nothing and certainly, at a minimum, should at least be tested within an acquisition program.

From these examples above, one can observe that intuition and judgment based on the practical experience or the "art" component are being used in building a solution. The science component is done when evaluating these techniques either through a live campaign which will either support or refute these "art" developed-type approaches.

Tip 7. Establishing Performance Benchmarks

In building predictive analytics solutions and, in fact, building any type of solution, it is always incumbent on the builder to demonstrate its performance. With more complex-type solutions that are more mathematically intense, it is even more paramount that the solution builder clearly demonstrates its performance. This does not mean that these solutions simply refer to the more normal statistical diagnostics as being the primary measure of performance. Although this dialogue has some merit within the business sector, this more purist approach towards mathematical results is going to garner far greater impact on the academic community. In business, however, performance needs to be measured in tangible results that can easily be measured and understood. For solutions within the consumer behavior area, this has much greater relevance Accordingly, we will look at a gains chart or decile table that has become the primary option in evaluating performance analytics solutions and, specifically, for consumer behavior-type problems. Let's review a typical gains chart example (Fig. 33.6).

This gains chart is produced by taking a predictive model, such as response, and applying it to a validation or holdout sample which has not been used to develop the model, but which is similar to the sample used for model development. The model is used to score the 200,000 customers from the validation sample, with decile 1 representing the highest scored names and decile 10 the lowest scored names. The response rate column represents the actual observed response rate in each decile, which occurred in a previous marketing campaign. If a model is performing very well, we should expect to see higher response rates in the lower deciles and lower response rates in the higher deciles. Rank-ordering of response rate is the actual model deliverable. Using this rank-ordering capability, we can then incorporate our business parameters regarding revenue per order and cost per marketing effort. This allows us to create that key business metric of ROI, where ROI is essentially an incremental measure that looks at the incremental revenue which is driven through incremental marketing costs. These results reveal how well the model differentiates on ROI. From this chart (Fig. 33.1), and purely from a profitable standpoint, our

% of Validation Sample	Validation Names	Response Rate	% of Total Responders	Response Rate Lift	Interval ROI	Modelling Benefits
0-10%	20000	3.50%	23%	233	145%	$26,667
10-20%	40000	3.00%	40%	200	75%	$40,000
20-30%	60000	2.75%	55%	183	58%	$50,000
30-40%	80000	2.50%	67%	167	22%	$53,333
40-50%	100000	2.25%	75%	150	-13%	$50,000
.						
.						
.						
90-100%	200000	1.50%	100%	100	-58%	$0

Fig. 33.6 Sample gains chart

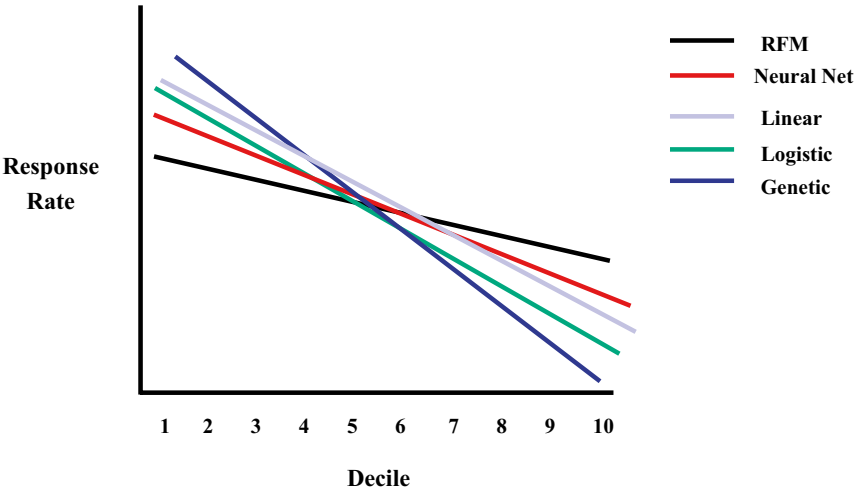


Fig. 33.7 Compering different modeling techniques

target group of customers would be the top 40%, which would support the ROI cutoff where, below 40%, the ROI becomes negative. If we promote 100% of the names here, then the modeling benefits are $0 as we are not taking any advantage of the model.

The actual modeling benefits are an opportunity cost that looks at how much incremental cost would be incurred to achieve the same level of performance but without a model. In this example, this is optimized at the top 40%, which is $53,333.

Once a model is created, this becomes the new benchmark in terms of performance. Newer approaches and techniques can always be measured against the initial model that was developed, as well as assessing whether the model's performance is eroding over time. This is done by again looking at the rank-ordering capability of the model when compared to other scenarios. Let's look at Fig. 33.7.

In Fig. 33.7, five different response models have been developed. A Lorenz curve or line, which plots the actual observed response rate at each decile, is created for each model. As you can observe, a line with more of a descending slope represents better rank ordering and, ultimately, a better model. Looking at these above results, we would surmise that the best model for response would be the one using the genetic algorithm approach.

In both using the model, as well as comparing model performances, we need to use metrics that are meaningful and tangible to the business. Looking at the statistical diagnostics is important in building models, but it is the business metrics which are easily understood that will determine the direction of predictive analytics solutions within a given business.

Tip 8. Interpreting Results Correctly

There is an often-used humorous phrase which attempts to encapsulate the sense that statistics can be misleading:: "There are lies, damned lies, and then statistics." In a way, there is some truth to this as numbers can always be interpreted in different manners. Furthermore, the numbers may point to other problems beyond just the generation of insights, as seen in Fig. 33.8.

In this mortgage insurance response model, we observe that the top 5% is capturing 80% of all the responders to this campaign. These kinds of results raise all kinds of red flags as typically excellent results might be that the top 5% accounts for 15%–20% of all the responders in that campaign, but definitely not **80%**. More probing and exploration of the data would ensue to better understand why these results exist. This might entail the analyst scrolling through the variable fields of the analytical file. Analysis of fields such as the dependent variable of purchasing mortgage insurance reveals that all these responders have previously been insurance buyers. This is impossible and, upon further investigation of the data as to what transpired, we discover that there was no preperiod which looked at behavior prior to the response event. Because of this failure to add the time dimension (pre- and post-periods) to the analytical file, all mortgage insurance responders would, of course, be now *previous* insurance buyers.

In this case. the results, rather than providing insights, instead led to questions of possible overstatement of performance. More importantly, however, it also required someone with the appropriate skillset to "actually get under the hood" of the data to identify what has caused this outcome and to resolve it accordingly.

As practitioners, we always pride ourselves on being able to explain a solution. This means that the actual variables or characteristics comprising a given solution should be easily described in terms of their impact on the predicted behavior. Correlation analysis reports and exploratory data analysis (EDA) reports can easily demonstrate the actual impact of a given variable on the desired behavior. However, multicollinearity, which looks at the interaction between variables, can cause model variables to switch in sign. For example, analysis (correlations and EDA) may find that age and income exhibit a positive impact on response, implying that older people and higher-income people are more likely to respond. When both variables become part of a multivariate

Fig. 33.8 Example of gains/decile chart demonstrating overstatement

% of Names Promoted	% of Mortgage Insurance Buyers
0-5%	80%
5-10%	5%
10-15%	5%
.....	10%
95-100%	1%

Resonpse Variable	Spending (last Yr)
Correlation Result	0.009
Confidence Level	10%

Spending Level	Response Rate
0 - 250	1%
251-500	2%
501-750	3%
750 - 20000	4%

Fig. 33.9 Example of outliers to recognize outliers in spending variable

solution, such as occurs with a logistic regression, we find that the actual sign for income is now negative, implying that lower-income people are more likely to respond. This switch in sign and impact to the modeled behavior by income needs to be understood in terms of what is causing it. Further analysis, such as correlation analysis between these variables (income and age), reveal that there is a very high correlation between age and income. How is this handled? Two courses of action can be taken, with the first looking at replacing income with another variable that maintains the performance level of the model but exhibits the same sign or impact with the modeled behavior that is observed in our other analytical reports (EDA and correlation). The second course of action would be to combine both age and income as one variable using techniques such as CHAID, whereby the output would detail the actual mechanics of how these variables should be combined.

In some cases, insights regarding certain characteristics may be overlooked since certain variables may exhibit no statistical impact due to outliers. Figure 33.9 lists an example of this situation.

Here the correlation results would indicate that spending has a minimal impact on response, yet the corresponding exploratory analysis report (EDA) would appear to illustrate a rather strong link between spending and response. The approach to handling this type of situation is to cap the outlier to a certain value (1000) using a mix of both automatic routines as well as some manual intervention in arriving at an appropriate value. Ignoring outliers, though, as seen from Fig. 33.9, can result in not being able to identify key variables that might comprise a predictive analytics solution.

The importance of this tip in interpreting results cannot be overemphasized. Solution builders within predictive analytics will strive for the best solution, which can sometimes have more of an academic focus on purity of results. This is particularly the case if new, inexperienced analysts are building these solutions. The resulting challenge for most organizations is to balance the mathematicians' and the practitioners' viewpoints. Without this dual perspective, in many cases predictive analytics solutions will not achieve their intended effect.

Relative vs. Absolute Results

In assessing results, it is often tempting to look at the actual numbers themselves, without considering them in some meaningful context. What do I mean by this? Figure 33.10 gives an example of two campaigns with test and control cell results.

If we simply considered the raw results, observing that there is a 1.5% difference between the test cell and the control cell in both campaigns, we might infer that the test cell in both campaigns is yielding the same impact. As you can observe, however, the test impact is far greater in Campaign 1 than Campaign 2. The real key to deriving the right insight from these numbers is to look at the figures in a relative manner as opposed to an absolute manner. This implies that the numbers be converted to lift measures. In the example above where the control cell is used as the benchmark, the test cell in Campaign 1 is yielding a lift of $200[(3/1.5) \times 100]$ versus a lift of $107[(23/21.5) \times 100]$ for Campaign 2. Looking at the numbers in the context of lift rather than absolute response rate numbers clearly provides better direction in terms of actionable next steps.

Another good example of using relative measures as opposed to absolute measures is when predictive models are used. The key metric in model evaluation, as indicated in past discussions, has been the rank-ordering of the observed behavior that we are trying to model. In effect, however, the model's real impact to the business is lift, or an increase in performance, as we are better able to target the right names. Figure 33.11 lists a decile or gains table, which demonstrates this capability.

In this gains table (Fig. 33.11), the observed or actual response rate which occurs within each decile is converted to response rate lift. The value of 233 means that the top decile responds at 2.33 times the response rate of the entire validation sample. In applying the response rate model to future initiatives, the decision on targeting names will be based on the relative measure of lift rather than the absolute measure of response rate. Rather than forecasting what the absolute response rate might be when using the model, practitioners are much more comfortable in determining what the lift in performance might be given the number of names which are selected by the model.

	Campaign 1	Campaign 2
Test cell	3% Resp. Rate	23% Resp. Rate
Control Cell	1.5% Resp. Rate	21.5% Resp. Rate
Difference	**1.5% Resp. Rate**	**1.5% Resp. Rate**

Fig. 33.10 Comparing relative to absolute results

% OF VALIDATION SAMPLE	VALIDATION NAMES	RESPONSE RATE	% OF TOTAL RESPONDERS	RESPONSE RATE LIFT	INTERVAL ROI	MODELLING BENEFITS
0-10%	20000	3.50%	23%	233	145%	$26,667
10-20%	40000	3.00%	40%	200	75%	$40,000
20-30%	60000	2.75%	55%	183	58%	$50,000
30-40%	80000	2.50%	67%	167	22%	$53,333
40-50%	100000	2.25%	75%	150	-13%	$50,000
.						
.						
.						
90-100%	200000	1.50%	100%	100	-58%	$0

Fig. 33.11 Gains chart/decile table

Action and Measure

In any course or seminar that I teach in data science or predictive analytics, one of my first comments refers to the absolute necessity of having a clear vision regarding how your solution or solutions will be implemented. Otherwise, the notion of "analysis paralysis" becomes the actual reality if there is no plan of implementation. Furthermore, any implementation plan should not have a long time horizon. Time horizons beyond six months in terms of implementation result in the solution having less relevance to the current business environment but also, more importantly, can lead to increased apathy amongst its key stakeholders regarding the implementation of the solution.

Presuming implementation is going to occur, several steps need to take place. First, we need to understand if the solution has been properly implemented. This is relevant when a solution is to be handed over to a third party. A good example of this is a model developed by the analytics area which is now going to be programmed into the data warehouse currently housed by the IT department.

Proper implementation of a solution requires a discipline of effective checks and controls. This means that the actual solution developer will score a current file for implementation while a third party (I/T) scores the same file. Both sets of scores should be identical. If the scores are different, then the programming logic used by the third party is different to the logic used by the solution developer. Investigation needs to occur in order to adjust this logic such that both sets of scores become identical.

At this point we have completed one check. Our checking is still incomplete, however, as we now need to compare the current score distribution to what we observed during model development. Figure 33.12 includes an example of what this might look like.

In this above table, we have two sets of model scores. Each row contains the decile (% of list), the minimum model score for each decile achieved when building the solution (validation sample), and the minimum model score for each decile as of the latest scoring run (current score). Ideally, we want both these distributions to look similar. From Fig. 33.12 below, you can see that this is not the case. Clearly, the current score distribution contains scores which are

% of List	Minimum Score (validation sample)	Minimum Score (current score)
0-10%	0.08	0.04
10-20%	0.07	0.03
20-30%	0.06	0.02
30-40%	0.05	0.01
40-50%	0.04	0.004
etc...		

Fig. 33.12 Comparing model predictions during time of development vs. current application

Age File	% of File (Development)	% of (Current Campaign)
< 30	20%	18%
30 - 40	20%	22%
40 - 50	20%	21%
50 - 60	20%	19%
60 +	20%	20%

Income	% of File (Development)	% of File (Current Campaign)
< 25 K	20%	20%
25 - 35 K	20%	19%
35 - 50 K	20%	17%
50 - 80 K	20%	23%
80 + K	20%	20%

Spend	% of File (Development)	% of File (Current Campaign)
< 25	20%	25%
25 - 50	20%	15%
50 - 75	20%	18%
75 - 100	20%	19%
100 +	20%	23%

Live in Quebec	% of File (Development)	% of File (Current Campaign)
Yes	20%	100%
No	80%	0%

Fig. 33.13 Comparing frequency distributions of model variables between time of development and current application

half of what was observed during model development. Once again, forensics are required to understand why this has occurred. The approach here is to examine each model variable and to look at the frequency distribution of the variable during both model development and the latest scoring run. If scores were comparable between development and the latest scoring run, then the frequency distributions of all model variables for both runs should also be very similar. However, in our example above, frequency distribution analysis of all model variables in both scoring runs were very similar, with the exception of the model variable "Living in Quebec", which had a negative impact on the overall score (Fig. 33.13).

In the example here, we found that 20% of our customers lived in Quebec during model development; during the recent scoring run, however, 100% of all customers lived in Quebec. With Quebec having a negative impact on the overall score, it makes sense that the scores would be much lower in the current run than had been observed during model development. This suggested to us that the model was only being applied to Quebec names. This was fine but, as solution providers, we needed to provide the right counsel in advising them on how to use the model in this type of context. We indicated that one course of

action might to develop one model for Quebec and another, separate one for the rest of Canada. Presuming that budgetary reasons do not allow for this redevelopment, we would use the model in its current form and adjust the lift expectations on how we expect the model to perform. Because the universe for implementation is very different than what was used to develop the model, this expected level of model performance would be adjusted downward. At the same time, we could also market a much smaller random sample representative of the universe used to develop the model in which Quebec comprised 20% of the records. Marketing's focus on targeting Quebec names exclusively for perhaps strategic reasons would still be achieved, while at the same time the model's performance can be properly assessed through the use of a random sample to Quebec names only. This would be the preferred approach as it would be representative of the desired business solution. At the same time, it would only take one to two days to redevelop the model since the analytical file has already been created. To redevelop the final model, we would simply extract the Quebec names and then rerun the appropriate statistics.

Effective implementation with all its checks and balances is critical to the success of predictive analytics as a key business process. However, the ability to determine if these solutions worked is critical to how appreciating how these tools will be used in future marketing initiatives. This implies that the establishment of a measurement framework be created to determine the effectiveness of these tools. Determining the right measurement framework is no easy task. Furthermore, the level of complexity in creating this framework can vary from industry to industry, as well as by channel. For example, the measurement of a bank's marketing program might involve several performance indicators. Since a bank can generate many different forms of revenue from a customer, such as ATM fees, services charges, interest revenue, and product revenue, the measurement of any program might want to look at how all these indicators are impacted within a certain campaign. Meanwhile, the measurement approach for a retailer might be as simple as looking at purchase behavior.

The digital world, and the ease in which the data is gathered in this environment, have simply augmented the demand for more effective measurement. At the same time, our increasing technologically-driven world has allowed the capability for more individualized tracking of behavior. This capability has allowed marketers and businesses to evaluate various initiatives and scenarios on a ROI basis, which is now based more on facts rather than assumptions.

To some extent, this appetite for ROI measurement has been further accelerated due to the financial meltdown of late 2008. As businesses began to recover, the need for accountability became more prevalent with increased demands on marketing departments to effectively measure all their activities. The growing impact of social media has not diminished this need; it has simply created new challenges in effectively measuring the impact of these types of programs.

Through this discussion, we can see the growing importance of measurement. But how does one go about it? Once again, as have seen in other areas

	Control	Test Piece 1	Test Piece 2	Do Not Promote
Modeled List	330,000 **Cell 1**	10,000 **Cell 2**	10,000 **Cell 3**	
Non-model List	40,000 **Cell 4**			10,000 **Cell 5**

Fig. 33.14 Example of campaign measurement matrix

of analytics, there is a discipline and approach towards effective marketing measurement. The first step is to understand what the business objectives are, as well as the learning objectives of the marketing initiative. The second step is to identify the key stakeholders within this initiative. In other words, which individuals are most accountable in terms of results due to this initiative? The third step is to identify what kind of data we will use, and how we will use it for measurement. This data component needs to be clear not only to the analyst building the measurement framework but also to the business people and stakeholders involved in this initiative. The last step involves the actual framework which, in many cases, can be a spreadsheet-type matrix. Figure 33.14 is an example of one such matrix.

In Fig. 33.14, we see a matrix that has been designed to answer several different questions in relation to a campaign, including:

- Did the campaign work?
- Did the model work?
- What communication strategy worked best?

Comparing cell 4 to cell 5 will yield results that determine if the campaign worked; comparing cell 1 to cell 4 yields results that determine if the model worked; and, finally, comparing cell 1 to cell 2 to cell 3 determines which communication strategy worked best.

Although the above measurement matrix represents just a simple example of how one might create the appropriate measurement framework, this final matrix is the culmination of a measurement process that is acceptable to all stakeholders.

The Do's and Don'ts in this section are a compilation of key practical concepts that I have experienced throughout my career as a data science. There are many others that I could write about but the ones in this chapter represent the most common ones that have arisen throughout my career. There is no academic aspect to these concepts except in some cases to challenge what the output means. As a data scientist, I have always been disciplined with the practical perspective of the discipline recognizing that the academic side provides tools which may or may not yield its anticipated benefit.