Fraud Detection and Risk Management in Internet Finance



Qingqing Ren • Hualing Liu

World Scientific

Fraud Detection and Risk Management in Internet Finance

Fraud Detection and Risk Management in Internet Finance

Qingqing Ren
Goldman Sachs & Co. LLC, USA

Hualing Liu
Shanghai University of International Business and Economics, China

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Control Number: 2026939414

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

FRAUD DETECTION AND RISK MANAGEMENT IN INTERNET FINANCE

Copyright © 2026 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 978-981-98-1845-7 (hardcover)
ISBN 978-981-98-1846-4 (ebook for institutions)
ISBN 978-981-98-1847-1 (ebook for individuals)

For any available supplementary material, please visit
https://www.worldscientific.com/worldscibooks/10.1142/14457#t=suppl

Desk Editors: Aanand Jayaraman/Cian Sacker Ooi/Kura Sunaina

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore

Preface

In the ever-evolving realm of Internet finance, the crucial role of robust risk management abilities cannot be overstated. For Internet finance companies, adeptly navigating challenges such as security risks, credit defaults, and information asymmetry is not just a necessity but a cornerstone for sustained growth and operational stability. Internet financial fraud, the focus of this book, not only usually takes various forms, making it difficult to detect and mitigate effectively, but also has the potential to disrupt and impact numerous steps in the workflows of Internet finance companies, magnifying the risks and exposing these platforms to significant economic losses. Consequently, anti-fraud research is not only imperative but also holds considerable significance for effective risk management within Internet finance.

This monograph rigorously explores Internet financial platforms as its primary research focus, aiming to thoroughly understand and control the inherent fraud risks embedded in their operations. The specific emphasis is placed on identifying and preventing both individual and group fraud. Providing a comprehensive overview, the book delineates the present state and challenges faced by the Internet finance industry and conducts research on risk management in Internet finance, centering around two major themes: credit scoring and fraud detection. The book studies Internet financial fraud employing advanced algorithms like machine learning, deep learning, and graphical models. Moreover, empirical research on risk management is conducted using diverse methodologies, including random forests, text mining, and semi-supervised adaptive models.

Existing research mainly relies on financial and operational data from platforms, yet most of these data have issues like lag and fraud. In contrast, this monograph leverages the abundant real-time user evaluation data existing in the Internet finance platforms. By constructing an Internet financial dictionary, the book identifies evaluation themes and text sentiment, subsequently developing a platform for public opinion

monitoring and an early warning system for timely and accurate identification of problematic platforms. Simultaneously, every single Internet financial platform serves as a research node to create an Internet financial platform network. Utilizing network analysis, complex network theory, and platform public opinion text data mining, the study constructs a correlation network model and a measurement model. This approach explores the impact of information dissemination on investment and lending activities within Internet finance platforms, alongside analyzing the correlation between information cross-propagation and platform risks as well as studying the mechanisms of risk contagion among different platforms. The monograph proposes viable risk mitigation plans in response to accumulated risks in existing platforms, providing practical guidance for policymakers to prevent and resolve Internet financial risks.

The core focus of the book is the construction of an anti-fraud model to identify individual and group fraud within the Internet financial field. The research further extends to building a comprehensive risk control system applicable to online credit platforms, ultimately enhancing the overall risk control abilities of the Internet finance industry. The integration of machine learning methods with financial risk management theory leads to a novel Internet financial risk measurement method. This method serves as a valuable reference for investors seeking to make informed and rational investment decisions.

The intended audience comprises university teachers and students, practitioners in the Internet finance industry, as well as regulators of the finance industry.

The financial system is the core of the modern economy, serving as a vital bridge for financing in the real economy. Despite ongoing improvements amid rapid economic development, the global challenge of unfair financial resource allocation persists. Addressing this, the International Monetary Fund introduced the concept of “inclusive finance,” centered on “microcredit” and “microfinance,” aiming to broaden the scope and depth of financial development.

The rise of user needs and Internet technology has given birth to Internet finance, a transformative force challenging traditional financial markets plagued by high costs and information asymmetry that easily leads to moral hazard. The development of big data, cloud computing, and mobile communications in recent years has facilitated the transformation of

traditional financial institutions and the entry of Internet companies into the financial industry, providing crucial technical support for the emergence of Internet finance. The emergence of Internet lending platforms, third-party payment, and credit review models has brought disruptive changes to the traditional financial markets.

Internet finance fosters a fair and orderly financial competition environment, particularly benefiting small and micro enterprises, which historically faced financing difficulties. With its low entry threshold, simplicity, and convenience, Internet finance promotes inclusive growth, offering unique advantages in microfinance sectors. Collaborating with traditional finance, Internet finance channels funds more effectively to the public and small enterprises, contributing to a fair, competitive environment and the diversified development of the financial system.

Furthermore, Internet finance enhances the efficiency of financial business by leveraging information technology to streamline transaction processes and reduce costs associated with information acquisition and processing. Its development requires fewer offline branches and employees, significantly reducing operating costs. Additionally, the application of artificial intelligence (AI) further improves the efficiency of financial services.

However, while infusing innovation into the financial industry, Internet finance has also posed challenges to supervision and regulation. Credit risks in Internet finance are exacerbated by the inherent contradictions of the Internet. While advanced information technology enhances transparency in transaction information and efficiency in payment and financing, the virtuality of the Internet increases the asymmetry of user information and capital flow information. Direct transactions may lead to moral hazard, as parties no longer need intermediaries. The imperfect mandatory information disclosure requirements in China further complicate evidence collection and investigations when security issues arise.

Being an information-based market with no or little time and geographical barrier to access funding, Internet finance operates on a scale and efficiency significantly superior to traditional finance. Consequently, the spillover effects of financial fluctuations are more pronounced. In the Internet financial market, especially the online lending market, most investors are individuals and they usually form groups of friends that exist solely on the Internet. As a result, they often time face limited access to

information. This limitation results in herd behavior when making investment decisions, causing a “butterfly effect” or a “trampled effect” in the market. Therefore, identifying fraudulent information in Internet finance and preventing risks from spreading are of utmost importance.

In summary, this book presents a cross-disciplinary research approach, combining financial economics, complex network analysis, and AI to address fraud identification issues in Internet finance. It explores applicants’ social network information in addition to basic customer data, enhancing user information accuracy. This book, integrating financial and economic theories with complex network methods and AI technology, studies the detection of Internet financial fraud and the mechanism of platform risk contagion to provide intelligent early warnings and decision-making consultation to prevent systemic risks in Internet finance. Qian Kejia, Guo Yiwei, Ding Yujie, Lin Bei, Zhao Chengyu, Xu Ziyan, and Jiang Ying completed the graphs in certain sections of this book.

About the Authors

Qingqing Ren is an experienced financial professional at Goldman Sachs in New York, USA, with a focus on residential credit and structured finance. She earned an MS in Management Science and Engineering from Columbia University and a BS in Finance from Peking University, where she graduated with honors from the Guanghua School of Management.

Beyond her thriving career in the financial industry, Qingqing is deeply passionate about the intersection of financial technology and data science. This passion extends to her academic pursuits, where she explores data-driven solutions to enhance risk management capabilities in the finance sector. Her primary research interests are in the areas of machine learning, online consumer credit, and financial fraud detection. Qingqing has made significant contributions to multiple academic papers exploring the application of machine learning and deep learning algorithms in credit risk assessment and fraud detection in the Internet finance space.

Qingqing's dedication to both professional and academic endeavors underscores her commitment to driving excellence in the dynamic landscape of finance.

Hualing Liu is a professor at the Shanghai University of International Business and Economics and Faculty Director of the Business Big Data Experiment Center. Professor Liu's primary research interests are in the areas of knowledge management, data-driven decision-making, and quantitative economics, and her recent research has focused on applications of management science and data mining in privacy protection and risk management in internet finance. Professor Liu earned her PhD in Management Science and Engineering from Donghua University and her BS in Applied Mathematics from Chongqing University.

Professor Liu has published over 50 papers in esteemed national and international journals and conferences, more than 20 of which have been indexed in EI Compendex. In addition to academic journals, she has

authored three academic monographs and holds one software patent along with six software copyrights. Her work has received multiple nominations and awards, with one of her monographs being named one of the “Top Influential Books in 2021” in the field of economics in the 2021 China Publishing Media Business Report.

In addition to her achievements in research, Professor Liu has been a dedicated educator for nearly four decades and is an advisor to master’s students.

Contents

Preface

About the Authors

Chapter 1 Introduction

1.1 Financial Risks and the Internet
1.2 Existing Research on Internet Finance and Current Research Challenges
1.3 Research Summary and Key Concept Definitions
1.4 Chapter Outlines

Chapter 2 Introduction to Internet Finance

2.1 Major Business Models of Internet Finance
2.2 The Development of Internet Finance
2.3 Internet Financial Technology
2.4 Advantages and Risks of Internet Finance

Chapter 3 Current Situation and Risks of Internet Finance in China

3.1 Current Status of Internet Finance
3.2 Current Status of Third-Party Payment
3.3 Development of P2P Online Lending Platform
3.4 Analysis of the Development of Crowdfunding
3.5 Current Status of Internet Wealth Management
3.6 Internet Financial Risks
3.7 Conclusion

Chapter 4 **Application of Machine Learning in Internet Financial Risk Research**

4.1 Current Status of Internet Financial Regulation
4.2 Internet Financial Risk Control in the Context of Big Data
4.3 Internet Financial Risk Control and Machine Learning
4.4 Application of Machine Learning in Credit Scoring
4.5 Application of Machine Learning in Fraud Detection Systems

Chapter 5 **Application of Deep Learning in Internet Financial Risk Research**

5.1 Background Introduction
5.2 Application of Artificial Intelligence in the Field of Internet Finance
5.3 Internet Financial Risk Control and Deep Learning
5.4 Application of Deep Learning in Credit Risk Control
5.5 Application of Deep Learning in Fraud Identification System

Chapter 6 **Application of Graph Model in Research on Internet Financial Risks**

6.1 Introduction
6.2 Graph Model Theory
6.3 Application of Graph Model in the Field of Internet Financial Risk Assessment

Chapter 7 **Random Forest-Based Risk Identification of Internet Financial Subjects**

7.1 Introduction
7.2 Random Forest Method
7.3 Kernel Regression Method

7.4 Feature Combination Method Based on Frequent Patterns
7.5 Fraud Identification Algorithm
7.6 Investment Recommendation Algorithm
7.7 Conclusion

Chapter 8 Risk Identification of Internet Financial Platform Based on Text Mining

8.1 Introduction
8.2 Text Mining Methods
8.3 Text Mining in Internet Financial Platform
8.4 Risk Identification of Internet Financial Platform
8.5 Fraud Identification on Online Lending Platforms
8.6 Fraud Detection for Online Loan Platforms
8.7 Conclusion

Chapter 9 Internet Financial Risk Assessment Based on Semi-Supervised Model

9.1 Background
9.2 Semi-Supervised Learning
9.3 Research Design
9.4 Simulation Experiment
9.5 Conclusion

Chapter 10 A Study of Risk Measurement of the Stock Index Futures Market Based on Quantile Regression Forests

10.1 Background
10.2 Model Theory
10.3 Empirical Research
10.4 Conclusion

Chapter 11 Research on VaR Risk Measurement Based on the EMD-QRF Model

11.1 Introduction
11.2 Research Method
11.3 VaR Risk Measurement Based on EMD-QRF
11.4 Empirical Study
11.5 Conclusion

Chapter 12 Summary and Outlook

12.1 Summary
12.2 Outlook

References

Index

Chapter 1

Introduction

1.1 Financial Risks and the Internet

The vitality of the macroeconomy is closely tied to a financial market of sufficient depth and stability. Major financial crises in history have often been accompanied by economic downturns. Recognizing this pattern, it becomes evident that financial risks demand continual attention and regulatory oversight.

The advent of network technologies, including cloud technology, big data, and mobile communications, has made the innovative "Internet + Finance" model popular. As an emerging financial industry, Internet finance is characterized by widespread participation, convenient payment methods, low operational costs, relatively transparent information, and abundant credit data, displaying notable inclusivity. However, the increasing popularity of Internet finance has also given rise to online fraud, causing significant societal harm and exposing regulatory gaps within this nascent industry. To foster the robust growth of Internet finance and align it with the needs of the real economy, the establishment of a comprehensive Internet financial fraud identification and financial platform risk mitigation system is imperative for maintaining financial stability.

In recent years, the rapid expansion of Internet finance has brought about the gradual emergence of systemic risks. Operating in a highly information-centric environment with minimal temporal and geographical constraints on fund circulation, Internet finance surpasses traditional finance in both scale and efficiency.[1] Notably, the amplified spillover effects of financial fluctuations pose challenges. In the Internet financial market, particularly the online loan sector, investors primarily consist of

individuals and groups of friends that form within the Internet ecosystem. Limited access to information prompts herd behavior during decision-making. Any conspicuous bias in the market can trigger a "butterfly effect" and herding behavior impacting other platforms and subsequently influencing the entire Internet financial landscape. Consequently, the identification of fraudulent information in Internet finance is crucial to prevent the spread of risks.

1.2 Existing Research on Internet Finance and Current Research Challenges

Existing research on Internet finance primarily falls into two main categories: the first category focuses on borrower behavior modeling and risk control of lending groups and individual borrowers on Internet financial platforms. This involves examining various borrower characteristics (such as loan interest rates, loan term, age, gender, credit, etc.) to assess whether timely repayments of loans will be made. The other category examines the supervision of the Internet financial sector, with a primary focus on legal aspects.

The research focus of this monograph lies in fraud identification and risk prevention within Internet finance. Most scholars conduct empirical research on fraud risks using public data from platforms like Prosper and Lending Club. For instance, Herrero-Lopez (2009) found through a Gaussian mixture model that borrowers in high-trust social groups can repay on time even in cases of higher risks.[2] Emekter *et al*. (2015) constructed a logistic regression model, and the results showed that higher FICO scores and credit ratings, lower revolving loan utilization, and lower debt-to-income ratios are associated with lower borrower default rates.[3] Building on this, Malekipirbazari and Aksakalli (2015) utilized random forest to evaluate borrower credit risk, proving that this method is superior to the credit evaluation methods of FICO and LC platforms to some extent.[4] Jin and Zhu (2015) innovatively categorized lending behavior into default, on the verge of default, and non-default types and compared the prediction efficacy of five models.[5] Guo *et al*. (2016) proposed using kernel regression method to quantify loan risk.[6] Dorfleitner *et al*. (2016) employed text

mining to analyze borrower loan descriptions in their applications and revealed the impact of positive descriptive words and spelling issues on loan approval.[7]

Currently, Chinese scholars primarily focus on the influencing factors of platform credit risk and how to use statistical methods suitable for big data to identify and give early warning of fraud risks in peer-to-peer ("P2P") online lending. However, the research on fraud risk identification mainly focuses on the individual level, with limited attention given to group fraud (including platform fraud). Data sources mainly include public data from platforms like Paipaidai and Renrendai. Zhang and Liu (2012) used panel data to study the rational herding effect in online lending, demonstrating that lenders adjust their credit assessments based on the decisions of other lenders and publicly available borrower information.[8] Wang and Liao (2014) discussed reducing credit risk through the P2P credit model and came to the conclusion that a combination of online and offline models is more effective in improving borrowing success rates.[9] Xiao *et al*. (2015) found through a ranked selection model that the borrower's personal characteristics, credit variables, historical performance, and borrowing information all have a significant impact on the credit risk of online loans.[10] Xia *et al*. (2015) improved the clustering support vector machine for default predictions.[11] Empirical evidence shows that this model can significantly improve the classification effect. Luo *et al*. (2017) proposed a "multi-level classification" method to detect "abnormal loan listings" on online lending platforms, effectively solving the problem of class imbalance in datasets.[12] A small number of scholars have adopted a unique approach, studying borrowers' credit risks based on their social relationships to determine the possibility of fraud. Freedman and Jin (2017) found that contrary to lender expectations, borrowers with more social ties are more likely to pay late or default.[13] The reason for this may be that there is a great deal of heterogeneity among group borrowers and some low quality borrowers may strategically present social ties to enhance loan appeal. Galak *et al*. (2011) explored investors' lending patterns based on Kiva data and social distance calculations and found that investors tend to lend to people similar to themselves.[14] Lin *et al*. (2013) used data from Prosper and observed the positive impact of online friendships on borrowing success rates, reducing interest rates and delinquency risks.[15] Wu and Wang (2013) built a credit

fraud group detection model for auction websites based on PageRank and social network analysis.[16] Wang (2017) studied the relationship between social networks, consumption behavior and fraud based on the social and consumption data provided by a company, thereby establishing a risk control model to predict user fraud behavior.[17]

In summary, research on P2P platform borrower credit evaluation is still in its early stages, and the application of data mining methods in this field needs further exploration. Logistic regression, decision trees, support vector machines, and random forests are commonly used credit risk assessment models. However, most studies use a single model, which may result in inaccurate and incomplete predictions. Additionally, research predominantly focuses on individual platforms, limiting the generalizability of results to other platforms. Moreover, fraud-related studies concentrate on the individual level, with no published literature addressing group collusion in fraud within the Internet financial sector.

1.3 Research Summary and Key Concept Definitions

1.3.1 *Research summary*

This monograph aims to leverage big data, artificial intelligence (AI), and cloud computing to construct profiles of Internet financial customers, study fraud identification and risk prevention in Internet finance, and build on existing research related to the herding effect of P2P online lending and social network dynamics. The research holds significant importance in ensuring all-weather and seamless financial risk control and management. Simultaneously, it aims to contribute significantly to the healthy development of Internet finance by promoting the integration of fundraising for small and micro enterprises through online platforms, thereby facilitating corporate innovation. The research will address the following key areas:

(1) **Internet Financial Fraud Risk Identification Based on User Behavior Analysis:** Utilizing financial big data, this approach involves showcasing user behavior from various angles, profiling Internet

financial customers, identifying fraud in Internet finance, and studying the propagation model of fraudulent effects.

(2) **Internet Financial Fraud Risk Identification Based on Social Relationships:** By mining key figures in fraud incidents within social networks, we are able to identify followers associated with these figures and thus label these users as likely to commit fraud. Thus, when multiple members within the same social network engage in online lending, we can analyze user behavior from multiple angles and assess the possibility of this social group conducting fraud, thus determining whether this is a group fraud event.

(3) **Integration of Multiple Data Mining Algorithms and Technologies:** This involves evaluating the credit profile of multiple Internet financial platforms from different dimensions and building effective early-warning systems for platform fraud incidents through the integration of various data mining algorithms.

This monograph, grounded in big data technology, applies complex network theories and related methodologies to the field of Internet finance, offering a novel solution to address fraud challenges in the field of Internet finance. The integration of machine learning with AI to construct a risk control model, particularly for the detection of Internet financial lending fraud, holds immense practical value, especially for real-time risk monitoring of Internet financial platforms.

1.3.2 *Key concept definitions*

1.3.2.1 *Definition of Internet finance*

Narrowly defined, Internet finance represents a new financial business model and refers to financial services conducted over the Internet, such as online banking, online securities trading, and online insurance. In essence, it involves the digital circulation of funds. As defined by Lizhen Liu (2014), Internet finance, in a narrow sense, comprises payment services and more sophisticated investment and financing businesses.[147]

In a broader sense, Internet finance refers to the Internetization of financial services. In addition to financial business in the narrow sense, Internet finance also includes areas like financial oversight and network security. The current landscape of Internet finance includes both traditional

financial institutions and non-financial institutions. Traditional financial institutions leverage Internet technology to conduct business and carry out financial innovations, while non-financial entities primarily include P2P online lending platforms, crowdfunding platforms, e-commerce companies, third-party payment platforms, and mobile wealth management services.

1.3.2.2 *Definition of the key concepts of Internet finance*

(1) Payment method

Payment stands as the foundation of the financial landscape, and within the domain of Internet finance, it is realized through mobile communication devices, that is, mobile payment.

The trajectory of mobile payment has witnessed an unprecedented surge, propelled by the rapid evolution of smart devices. According to the "Chinese Mobile Payment Industry Research Report in the First Half of 2019" by iiMedia Research, the transaction volume of mobile payments soared to 166.1 trillion yuan in the first half of 2019, marking a substantial increase from 27.74 billion yuan in 2018 and a 24.2% surge from the first half of 2018. With the continual expansion of mobile payment applications in daily life, it is anticipated that mobile payments will become even more prevalent. Simultaneously, the number of mobile payment users in China has continued growing and is expected to reach 790 million individuals in 2020, with an extended outreach to encompass a broader demographic, including elderly and child users.

The seamless integration of mobile payment with mobile communication networks, as well as its convergence with wired telephone networks, online banking, debit and credit cards, and other payment methods, enhances the convenience of mobile payments, allowing transactions to occur anytime and anywhere. Additionally, advancements in network security technology have broadened the scope of mobile payments, extending beyond routine daily transactions to facilitate substantial payments between enterprises. The advent of cloud computing has addressed the limitations of mobile communications, such as device storage capacity and computing speed, ensuring the efficiency of mobile payments by transferring storage and computing processes to cloud computing servers.

In China, mobile payment has experienced rapid development: individuals and institutions can establish accounts at the central bank payment center, enabling fully electronic payment and settlement processes. Mobile payment has substantially reduced reliance on cash circulation, and the settlement and payments of financial assets are now seamlessly conducted through the Internet. This transformative shift is reflected in the diminishing relevance of the second-level commercial bank account system.

(2) Information processing

Information serves as the foundation for allocating financial resources. For example, people can post their evaluations of bank services online, which helps others evaluate the bank's profitability and creditworthiness.

The key information lies in both borrower and lender data, and the borrower-related information is particularly important. Different financing models employ different information processing methods. Government regulation involves auditing and information disclosure mandates, requiring fund seekers to disclose accurate information. In contrast, securities firms, commercial banks, and credit rating agencies establish specialized entities to collect and sell information to lenders.

The difference in information processing methods makes the Internet financial model different from the traditional financing model. Under the Internet financial model, information about both parties to the loan is generated and disseminated through social networks and is standardized by search engines to form a continuously changing sequence of information. Cloud computing, integral to this process, ensures high-speed information processing capabilities, delivering a dynamic default probability for any borrower at a reduced cost.

Social networks, search engines, and cloud computing play different roles in information processing.

Social networks digitize real-life interpersonal relationships, providing a platform for voluntary information sharing. Human network behavior and the development of mobile communication devices form the basis of social networks. Numerous stakeholders, including individuals and institutions, hold operational and consumption information. While an individual stakeholder's data may be limited, social networks enable online information publication and aggregation, yielding comprehensive insights

into profit prospects and credit ratings, significantly reducing transaction costs and moral hazards.

Search engines enable users to extract information that best matches their needs from massive amounts of information, facilitating swift answers and recommending suitable individuals for queries. The integration of social networks and search engines enhances user efficiency.

Cloud computing, with its ability to efficiently process large amounts of information volumes through distributed computing tasks, addresses the substantial computing power demands of the financial industry, thus exerting a profound impact on the financial industry.

(3) Resource allocation

In Internet finance, resource allocation operates through a platform-based mechanism where both borrowers and lenders publish their demand online and then directly conduct transactions after automatic matching without going through intermediaries such as banks and securities companies. Take Lending Club in the United States as an example. Each loan in Lending Club provides borrower and loan information for investors to choose from. Lending Club assigns credit ratings to loan applications from A to G. The lower the credit rating, the higher the interest rate on the loan. The minimum subscription amount per order is US$25, which enables risk diversification. Lending Club provides investors with a trading platform and tools to build loan portfolios and handle possible defaults and delinquencies.

The rapid development of information technologies such as cloud computing, big data, and AI have enabled direct financial transactions between individuals which was a breakthrough that exceeded the traditional feasibility boundaries. Information between borrowers and lenders under the Internet finance model is almost completely symmetrical and highly transparent, which increases resource allocation efficiency and maximizes social welfare. Under this configuration, problems such as unregulated private lending and small and medium-sized enterprise financing can be well solved. Meanwhile, Internet finance platforms allow people or companies who do not know each other to form social network relationships, thereby enhancing cooperation opportunities.

1.3.2.3 *Definition of Internet financial fraud and anti-fraud concepts*

Internet finance, as a novel financial service model, caters to a substantial number of "long-tail" customers typically not covered by traditional finance. The greater uncertainty faced by this group of customers has exacerbated information asymmetry, leading to heightened risks like fraud and malicious breaches of contract. The rapid growth of Internet finance has unfortunately made it a prime target for fraudulent activities, evolving into an industrial chain that disrupts financial market stability and sparks broader social issues. In the realm of Internet lending, the establishment of trust relationships between two strangers through online platforms results in lending contracts. Yet, the concealment inherent in Internet transactions allows some borrowers, driven by the goal of acquiring assets, to have no intention to repay the loans from the beginning of the loan, causing losses for the platform. With ongoing advancements in Internet technology, various fraud methods have emerged, transforming individual fraud into organized, large-scale group fraud. This evolution poses formidable challenges to risk control in the financial industry.

As an emerging financial sector crucial to economic development, Internet finance's vulnerabilities in monitoring create fertile ground for fraudulent activities. Conventional anti-fraud models reliant on traditional rules are no longer effective in coping with the rapidly evolving landscape of fraud methods. Therefore, leveraging cloud computing and big data technology to analyze, calculate, and process Internet financial data for fraud identification and anti-fraud research holds significant academic value and broad application prospects.

Presently, research on Internet financial fraud identification and anti-fraud technology is in its infancy, with numerous challenges in applying existing machine learning and AI methods to anti-fraud efforts. Exploring and developing methods tailored for Internet fraud identification and constructing effective anti-fraud models carry substantial practical significance.

1.4 Chapter Outlines

This book takes Internet finance as its research object, focusing on the fraud risks existing in its operating model, including the identification and prevention of individual fraud and group fraud. This book is divided into 12 chapters, and the content is arranged as follows:

Chapter 1 introduces the background of the book's topic selection, discusses its research significance, defines important concepts, reviews relevant research and theoretical foundations.

Chapter 2 provides an overview of the historic development of Internet finance, examines the emergence and development of financial technology, and analyzes the advantages and disadvantages of Internet finance.

Chapter 3 analyzes the current situation and risk landscape of Internet finance in detail, providing a practical foundation for the main research work in the following chapters.

Chapters 4–6, respectively, explore the implementation of three methodologies — machine learning, deep learning, and graphical modeling — within the framework of Internet financial risk assessment models. These methodologies are predominantly applied across two major risk control themes: credit scoring and fraud detection. These three chapters establish both the theoretical and research foundations for the empirical research starting in Chapter 7.

Chapters 7–9 present comprehensive empirical research conducted on Internet financial risk control. The methodologies employed include random forests, text mining, semi-supervised adaptive models, and graphical models. The primary focus is on constructing an anti-fraud model designed to detect individual fraud and group fraud events in the Internet financial field. Subsequently, a fraud detection and risk control system is developed specifically for online credit platforms, contributing to the enhancement of the overall risk control standards within the Internet financial industry.

Chapters 10 and 11 combine machine learning methodologies with financial risk assessment theories to enhance methods for measuring the risks of financial products. This improvement provides valuable guidance for investors, facilitating informed and rational investment decisions.

Chapter 12 summarizes the research findings of this book and its main conclusions. It also offers a forward-looking prospective on potential further enhancements to the risk mitigation framework in Internet finance.

Chapter 2

Introduction to Internet Finance

2.1 Major Business Models of Internet Finance

The Internet ecosystem, centered around Internet technology and user value, marks a departure from industry boundaries established during the industrialization era. The traditional business ecological model is disrupted, resulting in increased diversity. The combination of "Internet + Finance" has created diverse forms of businesses, such as P2P online lending, crowdfunding, third-party payments, digital wealth management, digital financial institutions, and Internet financial portals.

2.1.1 *P2P online lending*

P2P online lending refers to direct online lending between individuals without the involvement of traditional financial intermediaries. Borrowers and lenders provide their information on the P2P platform, and upon approval, qualified borrowers can publish loan requests. Lenders then offer funds to those in need based on their risk-bearing capacity. The process eliminates the need for traditional financial middlemen. The concept originated in 1983 when Muhammad Yunus established Grameen Bank to provide loans to low-income individuals, serving as the prototype for P2P online lending. The first P2P lending platform, Zopa, emerged in London in 2005, leading to rapid P2P development in Europe and North America. China's P2P industry began with the founding of Paipaidai in 2007, followed by a surge in investor-backed P2P platforms. In its early stages, frequent borrower defaults resulted from a lack of communication between platforms and an imperfect citizen credit system. Post-2013, P2P platforms

experienced substantial growth, causing an imbalance between capital supply and demand, exposing vulnerabilities. The P2P industry, lacking proper supervision, became a diverse but mixed sector. Instances of malicious deception by P2P companies and platform defections occurred regularly.

2.1.2 *Crowdfunding*

Crowdfunding refers to the public solicitation of funds, where individuals or organizations seek project funding from the general public. The concept originated on the Kickstarter website in the United States, which established a platform for creative individuals to bring their dreams to reality through public fundraising. In 2011, China adopted the crowdfunding model with the introduction of the "Roll Call Time" website. In China, crowdfunding can be categorized into three primary forms: product crowdfunding, equity crowdfunding, and charity crowdfunding. Notable examples include Jingdong Crowdfunding, Taobao Crowdfunding, Modian (for product crowdfunding), Start-This is How You Revenge Mediocrity (for equity crowdfunding), Easy Crowdfunding, and Shuidi Crowdfunding (for charity crowdfunding). Unlike traditional financing methods, crowdfunding has a lower entry barrier and imposes no restrictions based on identity, gender, age, or occupation. Anyone with innovative ideas can initiate a project, resulting in a diverse array of projects. Ordinary individuals are free to invest in projects that align with their interests, such as design and invention, scientific research, artistic creation, entrepreneurial initiatives, and post-disaster reconstruction. Crowdfunding platforms facilitate communication among investors, allowing for collaborative decision-making and fostering a more informed investment environment.

2.1.3 *Third-party payment*

Established in 1999, Beijing Shouxin and Shanghai Huanxun stand as China's earliest third-party payment platforms. In its initial phase, the slow progress of e-commerce platforms hindered the growth of third-party payment platforms. This situation changed dramatically with the launch of Alipay by Alibaba in 2004. Fueled by the growth of Taobao, third-party payment platforms, led by Alipay, experienced rapid development. Within a

short span of 4 years, Alipay surpassed PayPal, securing its position as the world's largest third-party payment platform. Subsequently, platforms like WeChat Pay, online banking, and An Fu Tong emerged, further diversifying the landscape. Government-issued documents encouraged and guided the expansion of third-party payment services, leading to increased diversification and broadened business scopes, and third-party payment services entered a phase of stable growth.

While the growth rate of transaction volume in the third-party payment industry has moderated, it continues to be profitable. The implementation of various policies by the Chinese central bank has standardized third-party payments, mitigating adverse consequences from cutthroat competition. Simultaneously, it is foreseeable that the future development potential for third-party payment may shrink, emphasizing the significance of strategic transformation for its sustained growth.

2.1.4 *Digital wealth management*

Digital wealth management presents a novel approach that integrates conventional wealth management with the Internet, utilizing the online space as a sales channel for financial products. This method allows investors to directly purchase financial products on the Internet without visiting a bank or a fund. Digital wealth management addresses the limitations of traditional financial institutions in small and microfinancing. It boasts a low threshold for wealth management, enabling the public to achieve relatively high returns on dispersed funds, surpassing the returns from standard bank savings. Moreover, it offers high liquidity, allowing quick redemption within hours or, at most, two days. The presence of mature third-party payment platforms further simplifies the process of engaging in digital asset management, removing the constraints of time and location for investments. Well-known examples in China include Yu'e Bao and Li Cai Tong.

2.1.5 *Digital financial institutions*

Digital financial institutions refer to banks, securities firms, and other financial institutions that use information technology to transform or reconstruct traditional operating processes and realize comprehensive electronic operations and management.

The infusion of information technology proves effective in significantly reducing the operational costs of financial institutions. It also addresses information asymmetry issues, mitigates risks, and enhances their capacity to serve the real economy. Moreover, banks can proactively meet the challenges posed by Internet finance through the establishment of an integrated "online banking + e-commerce + financial supermarket" platform.

Under the Internet financial model, traditional financial institutions can offer increased convenience to the public. For instance, individuals no longer need to visit the bank for deposits and withdrawals, and stock transactions can be conducted without the necessity of visiting business offices. Looking ahead, traditional financial institutions need to strategize how to capitalize on their strengths, such as robust balance sheets and comprehensive risk control systems, to effectively harness various information technologies, thereby bolstering their competitiveness in the market.

2.1.6 *Internet financial portal*

Internet financial portals leverage the efficiency, timeliness, and comprehensiveness of Internet information dissemination by employing a vertical price comparison method to showcase diverse financial products online. Users have the opportunity to compare various products before finalizing transactions. This model entails low policy risk, as these portals function solely as platforms for advice and price comparison without involvement in the sales process or the flow of funds.

The emergence of Internet financial portals has shifted users from an information disadvantage position prevalent in the traditional financial model. This transformation has effectively mitigated the information asymmetry problem between users and financial institutions.

2.2 The Development of Internet Finance

The inception of Internet finance dates back to the establishment of the Security First Network Bank in the United States in 1995, paving the way for its subsequent development across Europe, America, and Asia. As the economics in various countries improved, traditional financial services transitioned online, and the rapid rise of Internet finance significantly impacted both traditional finance and the global financial landscape. The development of Internet finance in China can be delineated into three distinct stages:

(1) **First stage (before 2005):** During this stage, traditional financial institutions embraced Internet technology for their operations, with major commercial banks introducing online banking services. For instance, in 1997, China Merchants Bank launched its online banking platform. Simultaneously, the emergence of Taobao and Alipay introduced a new business model — electronic commerce — signifying the advent of the Internet finance era in China.

(2) **Second stage (2005–2012):** This period witnessed the emergence of further developed models such as crowdfunding and online loan platforms, catalyzing the rapid growth of Internet finance. The rise of third-party payment platforms, notably Alipay, addressed credit issues in e-commerce. Advances in mobile communication technology expanded the reach of third-party payment to various services. In 2007, the establishment of China's first P2P online lending platform, Paipaidai, marked a new phase in the development of Internet finance.

(3) **Third stage (2013–present):** The emergence of Yu'e Bao has made 2013 known as the "First Year of Internet Finance" in China. Internet finance entered a phase of rapid expansion, with third-party payment platforms maturing, crowdfunding platforms addressing challenges across various sectors, and P2P lending platforms experiencing a boom in development. However, beneath the swift economic progress, risks accumulated, evidenced by frequent defaults and collapses of P2P online loan platforms. Simultaneously, student loans on the Internet finance platforms garnered public attention, drawing increased scrutiny from regulatory authorities.

This section will delve into a detailed analysis of the development of Internet finance, focusing on online lending, fundraising, payment, and digital wealth management.

2.2.1 *Online lending*

P2P online lending, as a primary form of online lending, has consistently drawn significant attention. As the name implies, P2P online lending involves direct lending between individuals through Internet platforms. In 1983, Muhammad Yunus, hailed as the "father of microfinance," established Grameen Bank to specifically provide loans to impoverished individuals, laying the groundwork for the P2P network lending prototype. The world's earliest P2P online lending company, Zopa, originated in London, England, in 2005. This model later expanded to other countries, including the United States and Germany, with the emergence of Prosper and Lending Club in 2006, introducing innovative lending models. The inception of China's P2P online lending can be traced back to the establishment of Paipaidai in 2007, marking a rapid development in China's P2P industry, leading globally in the number of loans and investors.

P2P online platforms in China can be categorized into two types: fully online platforms and those operating on the O2O model, where borrowers undergo offline review. Based on distinct asset types, these platforms can further be classified into those providing project financing loans to small and micro enterprises and those offering personal loans.

As P2P online lending experiences rapid expansion, its operational strategies evolve. The initial focus of platforms like Paipaidai and Renrendai was primarily on the personal credit market, employing a single lending model based entirely on online transactions. However, since Renrendai and Youzhong Credit Industry jointly launched a new O2O credit model in December 2012, Renrendai has seen swift development. Subsequent P2P online lending platforms, such as Lufax and Hongling Venture Capital, predominantly focus on financial asset transactions and corporate project financing, showcasing a diversified trend in the P2P online lending market.

The P2P business model in China introduces risks such as insufficient supervision of intermediary accounts, potential product alienation, illegal fundraising, and other associated risks. Different business models exhibit

unique risk characteristics, presenting challenges for effective supervision. In 2018, numerous platforms in the P2P industry witnessed large-scale defaults, revealing escalating credit risk issues and leading to a wave of platform closures. Many investors incurred losses, diminishing the competitiveness of the overall financial market and prompting the urgency of regulatory measures for P2P platforms. However, the substantial participation in P2P platforms and the high flexibility in fund acquisition make it increasingly challenging for the regulators to supervise Internet finance effectively.

2.2.2 *Internet crowdfunding*

Crowdfunding, as the name suggests, involves individuals or organizations raising funds from the public. The American website Kickstarter is the earliest form of crowdfunding. The website platform allows creative people to raise funds to make their dreams come true. There are three main forms of crowdfunding in China: product crowdfunding, equity crowdfunding, and charity crowdfunding.

In product crowdfunding, investors invest in the development of a product or service, but what they are actually buying is a product that has not yet been produced. The products mainly focus on scientific and technological products, literary and artistic products, etc. In recent years, this form of Internet finance has experienced exponential growth. For example, the scale of investment funds in cultural projects has increased significantly. In 2011, there were less than 10 cultural projects crowdfunded in China, but in just 2 years, the number of cultural projects participating in crowdfunding has reached 150. The government's increasing support for "entrepreneurship and innovation" has promoted the rapid development of cultural crowdfunding.

The prosperity of product crowdfunding is due to its novel model. This model enhances user stickiness and has great appeal to investors who love innovation and creation. The low investment threshold expands the investor base, while financiers can better integrate resources to improve the possibility of project completion.

Equity crowdfunding has been controversial since its inception. Equity crowdfunding uses project equity as returns for its investors. It is a new financing model that allows investors to receive equity according to the

percentage of investment. It pools funds from multiple individual investors to support the fundraiser's business operations or other social ventures. The Company Law stipulates that the number of shareholders of a joint stock company may not exceed 200, and that of a limited partnership may not exceed 50. At the same time, the Securities Law stipulates that any issuance of securities to unspecified entities or to more than 200 specified entities shall be approved by the Securities Regulatory Commission or the competent department under the State Council, and transactions shall comply with the relevant laws and regulations. If these two regulations are violated, the equity crowdfunding will be deemed as illegal fundraising. Therefore, to avoid such risks, equity crowdfunding adopts an "online + offline" model where projects are presented online and fundraising is done offline.

Equity crowdfunding is highly regulated due to its characteristics. The Private Equity Crowdfunding Financing Management Measures (Trial) (Draft for Comments), passed in 2014, defines the nature of platforms and investors. The platform, as an intermediary, provides information and fund transfer assistance to both investors and financing parties. Investors must provide proof of assets and be able to bear certain risks. The "Notice on Special Inspection of Institutions Carrying out Equity Financing Activities Through the Internet," issued by the Securities Regulatory Commission in 2015, clearly stipulates the "transparent, small-amount, and non-institutional" characteristics of equity crowdfunding. Without approval, no entity or individual is allowed to engage in equity crowdfunding, further demonstrating the Chinese government's determination to strengthen the supervision of equity crowdfunding.

In public welfare crowdfunding, investors' funds are mainly used for certain public welfare enterprises. Investors do not expect specific returns and pay more attention to the participation process. For example, Qschou was founded in 2014. People seeking help for critical illnesses can set up aid projects on the platform. The platform can gather the power of the public and collect donations to help those in need. As of 2018, the amount of donations raised through Qschou has exceeded 20 billion. The public welfare crowdfunding platform allows every ordinary person to participate in the development of China's public welfare undertakings.

2.2.3 *Internet payment*

Third-party payment platforms are independent organizations that facilitate transactions between two parties by providing guarantees to both parties. They only play a transit role and do not have ownership of the funds. In 1999, the emergence of the PayEase marked the start of China's third-party payment, and the development of China's third-party payment platform has so far passed through three major milestones.

Alipay was born in 2003, and Yu'e Bao's scale exceeded 250 billion yuan in 2004. The third-party payment platform began to grow rapidly and then went through a decade of extensive development. However, due to the immaturity of relevant laws and regulations and inadequate supervision, platform problems arose, and bad behaviors such as illegal credit card redemption, illegal money laundering, and illegal embezzlement of funds appeared one after another.

Accordingly, in 2010, the People's Bank of China issued the "Measures for the Administration of Payment Services of Non-Financial Institutions," which stipulates that non-financial institutions must obtain a license in accordance with the law to become a payment institution. Without a license, no non-financial institution or individual may engage in payment business or engage in payment business in a disguised form. After the promulgation of this measure, the People's Bank of China issued the first batch of licenses to 27 companies, including UnionPay and Alipay. The third-party payment industry continued to develop rapidly during this period. The market size of third-party payment reached 1.6 billion yuan in 2013, with UnionPay, Alipay, and Tenpay accounting for the largest market shares. At the same time, independent third-party payment enterprises are maturing. Companies represented by UnionPay have mature offline transaction networks and have an absolute advantage in offline transaction scale. Internet-based payment platforms led by Alipay are far ahead in online transaction volume. An independent third-party payment company that does not rely on the group's resource advantages is different from the above two models. It does not rely on e-commerce websites, does not assume guarantee responsibility, and only provides payment service solutions. Independent third-party payment services led by Kuaiqian have tapped into emerging markets such as finance and aviation and achieved breakthroughs in their respective fields.

The emergence of mobile payment has further promoted the development of third-party payment. The well-known Alipay and WeChat Pay are typical representatives of mobile payment. The popularization of communication technology, the improvement of mobile payment functions, and the rapid growth of mobile payment market demand have all contributed to the rapid growth of third-party mobile payment.

In 2015, the People's Bank of China issued the "most stringent" regulatory regulation in history, the "Measures for the Management of Online Payment Business of Non-Bank Payment Institutions (Draft for Comment)," which imposes limit on the amount of payments with personal account balances. The measures have a great impact on third-party payment platforms based on e-commerce. At the same time, payment institutions are not allowed to open payment accounts for institutions engaged in financial business, causing third-party payment companies to readjust their business scope and suffer losses due to the cessation of some businesses. The measures also stipulate that third-party institutions are not allowed to support clearing, which aims to maintain the security of the payment system and reduce systemic risks.

The 2019 Q1 China Mobile Payment Market Research Report released by iiMedia Research pointed out that the application scenarios of third-party payment platforms continue to expand, mobile payments have penetrated into users' main life scenarios, and the frequency of transactions and the overall scale of transactions have increased rapidly. In 2018, China's mobile payment transaction volume reached 277.4 trillion yuan, and in the first quarter of 2019 alone, the transaction volume reached 83.9 trillion yuan.

2.2.4 *Internet wealth management*

Internet wealth management is another innovative model of Internet finance. The emergence of Yu'e Bao marks the beginning of Internet wealth management in China. Individual users deposit money into Yu'e Bao, and the funds are used to purchase shares of money market funds, which mainly invest in short-term monetary instruments such as treasuries and certificates of deposit. Therefore, the investment has low risk and generates stable returns, meeting the public's risk-free investment needs.

About a year after the launch of Yu'e Bao, the number of users reached 100 million, and the assets under management exceeded 500 billion yuan. With the successful case of Yu'e Bao, other Internet financial products such as Lingqianbao and Huoqibao are also becoming increasingly popular. In the early days of the development of Internet wealth management, when bank deposit rates were too low and the public's wealth was shrinking, the relatively high annualized returns of Yu'e Bao and other similar products enabled them to attract a large number of users and experience explosive growth, satisfying the strong investment demands from individual investors. At the same time, these financial products have ushered in an era of fragmented wealth management, with an extremely low investment threshold, simple purchase and redemption rules, and easy accessibility for users to purchase and redeem anytime and anywhere. This new wealth management model meets consumers' needs and reduces users' psychological burden, so the market size of financial products such as Yu'e Bao continues to expand.

Subsequently, Internet finance entered a stable development period, with annualized returns gradually declining and user accumulation growth slowing. As for Yu'e Bao, since 2014, the annualized rate of return has continued to decline, and the market size has gradually shrunk. Relevant research shows that the decline in the annualized rate of return of Internet wealth management is related to the continuous easing of liquidity in the banking system and the interbank lending rate.

Money market funds mainly invest in short-term monetary instruments such as central bank bills, interbank loans, and government bonds. Several interest rate cuts in China have led to a decline in the risk-free interest rate, which in turn affects the annualized return of Internet wealth management products. At the same time, the country's continued monetary easing has led to a decline in the demand for money market funds, and the four major banks do not accept deposits for these wealth management products, which has also contributed to the decline in the interest rates of wealth management products.

Although the annualized yield of Internet wealth management products is gradually declining, their competitiveness remains undiminished. The low investment threshold and T+0 settlement have advantages that traditional wealth management products do not have. Therefore, major traditional financial institutions have also started to launch their own

Internet wealth management products to lower the investment threshold. Since they do not have online payment functions, their funds are safer than products offered on platforms such as Yu'e Bao, so they are preferred by more and more investors.

2.3 Internet Financial Technology

Fintech refers to the application of diverse scientific and technological advancements to innovate products and services within the traditional financial industry, thereby improving efficiency and substantially reducing operating costs. Its inception aligns with the evolution of cloud computing, big data, and mobile communication technologies, catalyzing the progress and ingenuity of Internet finance.

Shusong Ba, at the 2017 Asian Financial Forum, delineated the evolution of financial technology into three distinct stages. Fintech 1.0 represents a phase where IT companies indirectly engage in financial activities, while traditional financial institutions leverage IT technology to digitize their operations and enhance efficiency — exemplified by the credit and clearing systems of traditional financial institutions. Fintech 2.0 signifies a transformation of traditional financial channels, as the industry establishes online business platforms, aggregates data from a multitude of users, facilitates information sharing, and integrates asset management, payments, and transactions. P2P online lending, Internet funds, and insurance epitomize this stage. Fintech 3.0 denotes the utilization of cutting-edge IT technologies, such as cloud computing, big data, artificial intelligence, and blockchain, risk pricing model updates, information collection processes, investment decision-making, and other facets of traditional finance. This stage is marked by advancements in supply chain finance, intelligent investment advisory, and similar innovations, effectively enhancing the overall efficiency of traditional financial practices.

2.3.1 *Cloud computing*

Cloud computing was first proposed at the 2006 Search Engine Conference. It is a process that combines multiple computing resources and provides them quickly through automated software operations, requiring minimal human intervention. There are three main business models for cloud computing: Amazon offers computing resource leasing, Google provides online application software services, and IBM sells cloud computing servers.

Cloud computing has significant economic value. Cloud computing enables enterprises with service demand at different time intervals to cross-configure service times, improving server usage efficiency. Enterprises can purchase servers according to their own needs, reducing operating costs. Additionally, cloud computing can help companies with data batch analysis needs to improve analytical efficiency.

The various benefits of cloud computing have attracted the attention of traditional financial institutions. Many large banks have adopted cloud computing technology to aid in their development, reduce risks and costs, enhance data analysis capabilities, and shorten response time. However, for many small traditional financial institutions, the high cost of using cloud computing is prohibitive.

2.3.2 *Big data*

Big data refers to large and complex data sets that traditional data processing software cannot handle. The concept of big data first appeared in a 1998 USENIX meeting by the chief scientist of SGI. The report highlights that the rapid growth of data volume makes it difficult to understand, obtain, process, and organize data. Big data is a term created to describe this problem and to stimulate thinking in computing.

Since 2014, relevant technologies, products, and standards have continued to develop, and the big data ecosystem has gradually improved. Big data provides new ways of thinking to understand complex systems and new means to explore objective laws and transform nature and society. Big data has not only brought about the rise of Internet finance, but its powerful analysis technology has also penetrated into all aspects of society. Search engines enable querying of large data sets, providing a foundation for

Internet finance. Traditional financial institutions can utilize big data technology to offer investors dynamic default probabilities of borrowers, reducing information asymmetry and risk. Big data also plays a vital role in promoting legal progress and optimizing government functions.

2.3.3 *Mobile payment*

Mobile payment is the process by which consumers use mobile devices to send instructions to banks and financial companies to transfer funds. This method allows consumers to conduct transactions anytime and anywhere without being limited by geographical restrictions or business hours. It offers the advantages of convenience, 24/7 availability, and eliminates the need for exchanging change.

The popularity of mobile devices has been greatly promoted by the development of communication equipment and technology, which has also led to the rapid growth of mobile payments. In 2010, the People's Bank of China issued the "Measures for the Administration of Payment Services of Non-Financial Institutions," which clarified the operating thresholds for mobile payment companies. According to the law, mobile payments should be subject to the supervision of the People's Bank of China. Since 2013, mobile payments have experienced explosive growth, and the number of users has increased exponentially. Research reports indicate that in 2019, Alipay and WeChat dominated the third-party payment market in China.

The development of mobile payments has fundamentally transformed the traditional payment model. The rapid development of Internet technology has enabled mobile payments to provide new momentum for China's economic development, improvement of people's livelihoods, and social progress.

2.4 Advantages and Risks of Internet Finance

2.4.1 *Advantages of Internet finance*

The emergence of Internet finance has diminished the role of traditional financial institutions in activities such as payment and financing. The development of Internet information technology has led to an increase in online payments, which has weakened the payment function of banks to some extent. The emergence of third-party financial institutions and payments has solidified the role of Internet finance in the payment function, breaking rigid payment structures and demonstrating innovation and upgrading of payment platforms. Traditional commercial banks are also facing challenges in financing services. From an income perspective, the investment return rate of Internet finance is higher than that of bank deposits, making it more attractive to investors. The benefits of Internet financing, including low mortgage thresholds, fast approval processes, and low transaction costs, make it an attractive option for small and micro enterprises struggling to secure financing.

Internet finance models eliminate the need for intermediaries such as banks, exchanges, and securities firms, streamlining the supply and demand of funds. Direct transactions can occur more conveniently which significantly reduces transaction costs. This model fully relies on the market, weakens the role of financial intermediaries, and achieves the same allocation efficiency as indirect financing, approaching a state without financial intermediaries. However, it downplays the professionalism of the financial industry and lowers the entry barriers for financial participation. Ordinary market participants can conduct various financial transactions through the Internet, thanks to simpler and more user-friendly operations. This model also alleviates the problem of financing difficulties for small and medium-sized enterprises, producing social benefits.

2.4.2 *Risks of Internet finance*

Although Internet finance has brought innovation to the financial industry, it has also created regulatory challenges. While it can significantly reduce transaction costs compared to traditional finance and has advantages such as convenient payment and inclusive benefits, insufficient supervision of P2P platforms has led to misappropriation of funds, absconding with money, and bankruptcy. It is clear that there is a need for better regulation

and oversight in this area. At the same time, the presence of information asymmetry between borrowers and lenders can easily lead to moral hazard and adverse selection problems, resulting in increasingly prominent credit issues and frequent borrower default risks.

Credit risks in Internet finance arise from the inherent contradictions of the Internet. The use of advanced information technology can make transaction information more transparent, reduce information asymmetry, and improve payment and financing efficiency. However, the Internet's virtual nature also increases the degree of information asymmetry in user data and capital flows. Direct transactions between parties may cause moral hazard, as intermediaries are no longer necessary. At the same time, the imperfect mandatory information disclosure mechanism and users' weak risk management capabilities in China make it difficult to obtain evidence and review when issues arise.

2.4.3 *Impact of Internet finance on China's financial landscape and regulatory considerations*

In recent years, China's economy has experienced a downturn. The rising default risk has put pressure on bank assets, leading to an increase in non-performing assets and a decline in bank profits. The non-performing loan ratio of commercial banks has also increased, contributing to the decline in profit margins. Additionally, the development of Internet finance has significantly facilitated the marketization of interest rates in the country. Interest rate marketization refers to the shift from government control to market control, where the cost of funds in different industries is determined by market supply and demand. The rise of Internet finance has led to competition between traditional banking institutions and Internet finance, promoting the marketization of cost of funds. Third-party institutions' participation reduces financing costs, promotes financial market innovation, and improves efficiency.

Currently, China's financial system is dominated by indirect financing and the banking system, with direct financing and capital markets as supplements. Therefore, the credit system in China is not yet sound, interest rates are controlled, and the consumer protection mechanism is not perfect. Additionally, funding supply and allocation are limited, and there is a lack of investment channels for the public. Internet finance, on the other hand, is

easy to operate, has low investment thresholds, and is highly transparent. Internet finance offers a wide range of products at low prices, with high efficiency and strong innovation, meeting the diverse investment and financing needs of individuals and enterprises. The risk assessment mechanism of Internet finance provides users with accurate and personalized services, promoting the innovation and transformation of traditional financial institutions.

In summary, Internet finance, as a new player in the financial industry, has significant impacts on the traditional financial sector. It has helped to enhance the Chinese financial market, promote innovation in financial products, and provide a more efficient and convenient market environment. However, it is important to address regulatory and security concerns that may arise from Internet finance.

Chapter 3

Current Situation and Risks of Internet Finance in China

3.1 Current Status of Internet Finance

With the deep integration of Internet technology and the traditional financial industry, not only has the concept of financial technology been born, but also four main forms of Internet finance have gradually formed: third-party payment, P2P online lending, crowdfunding, and Internet wealth management. Fintech is promoting the upgrading and transformation of China's Internet finance.

3.2 Current Status of Third-Party Payment

Since 2010, the transaction volume of China's third-party payment market has grown rapidly at an average annual rate of over 50%, making it a global leader. According to iResearch's "Q3 2019 China Third-Party Payment Quarterly Data Release," China's third-party mobile payment transaction volume reached 56 trillion yuan in Q3 2019. The proportion of mobile consumption in third-party mobile payment transactions increased from 21.4% in Q2 2019 to 22.2%. Additionally, offline QR code payments are gradually expanding to remote areas and third-tier cities, bringing in new users and payment volumes are expected to grow considerably in the future.

Since the first quarter of 2018, the growth rate of third-party payment transactions has continued to slow down. This indicates that China's mobile payment market has ended its rapid growth period and entered a stage of

steady growth. According to Figures 3.1, 3.2, and 3.4, it is evident that the offline QR code payment market is growing at a faster rate than the overall market. This growth is driving the mobile consumption sector to increase its share. The market share of the personal application sector has declined given it has entered a more mature stage with modest growth. According to Figure 3.3, Alipay and Tenpay hold a monopoly position with market shares of 54.5% and 39.5%, respectively. Figure 3.4 shows that in the third quarter of 2019, the offline QR code payment transaction volume was 8.6 trillion yuan, with a quarter-on-quarter growth rate of 7.8%. Additionally, Figure 3.5 indicates that the NFC payment transaction volume on mobile smart terminals reached 91.8 trillion yuan, with a month-on-month growth rate of 9.8%. In Q3 2019, third-party Internet payment transactions in China amounted to approximately 6.1 trillion yuan, which decreased compared to the same period in 2018. According to iResearch, this decline is due to the regulatory "three reductions" requirements. As a result, the volume of online lending transactions has continued to decline, and the growth of Internet financial business transaction volume has been greatly restricted. As showed in Figure 3.6, the third party Internet payment system market share was dominated by Alipay (35.7%), followed by UnionPay (19.0%) and Tenpay (13.7%) as of Q3 2019.

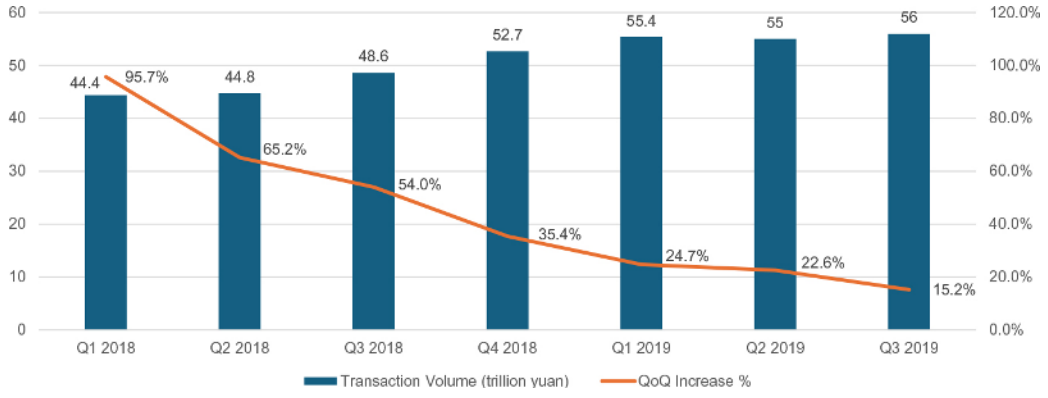


Figure 3.1. China's third-party mobile payment transaction size from Q1 2018 to Q3 2019.

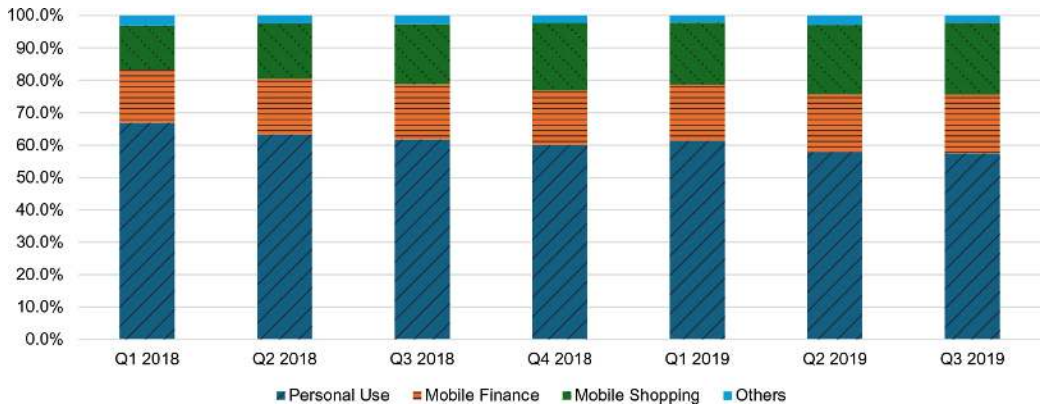


Figure 3.2. China's third-party mobile payment transaction size by application from Q1 2018 to Q3 2019.

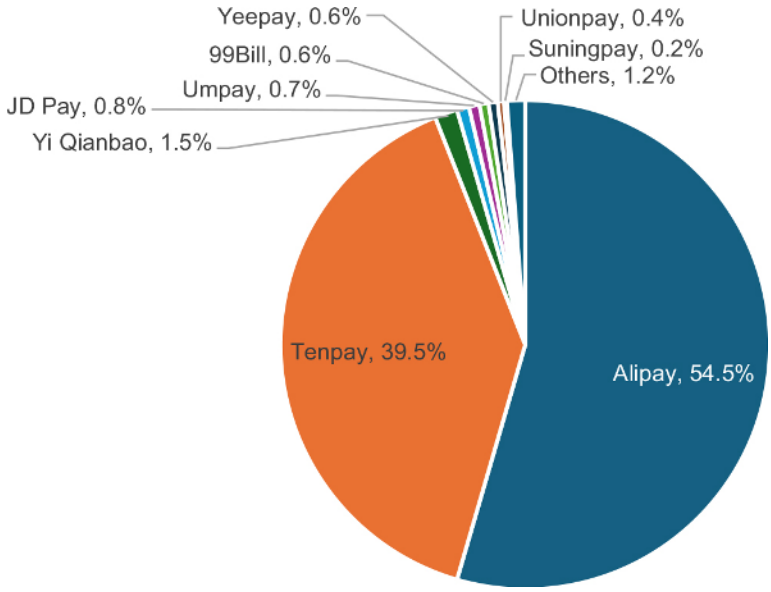


Figure 3.3. Market Share of China's third-party mobile payment transaction volume in Q3 2019.

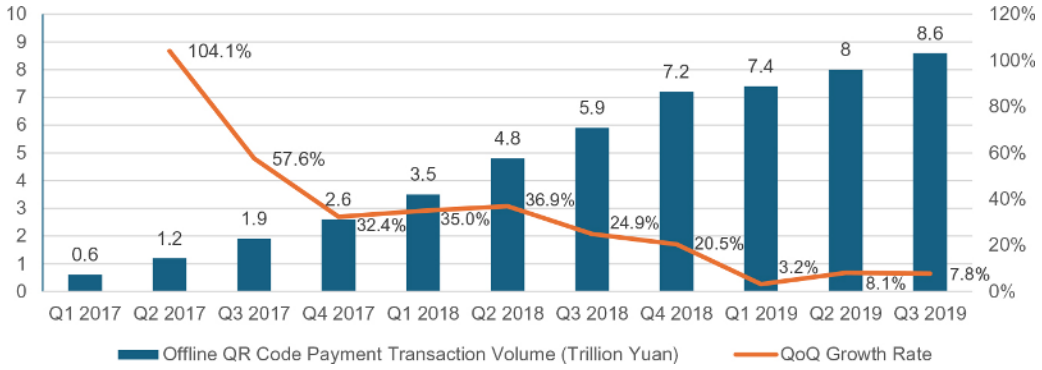


Figure 3.4. Offline QR code payment transaction size in China from Q1 2017 to Q3 2019.

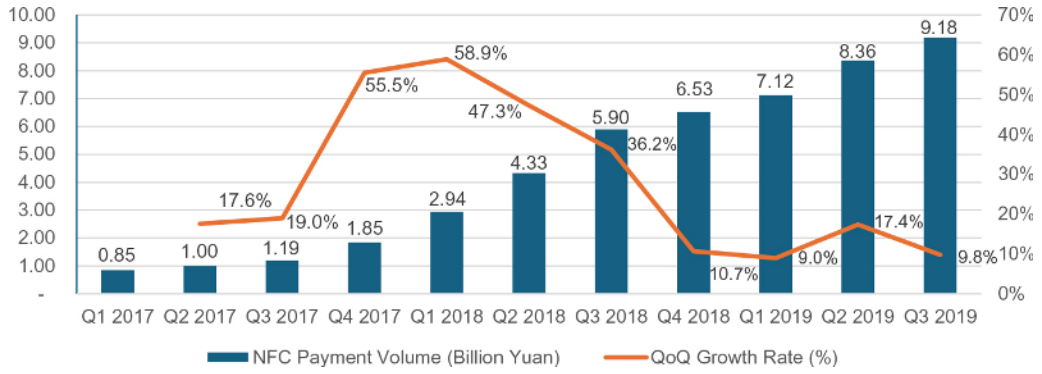


Figure 3.5. China's NFC mobile payment transaction size from Q1 2017 to Q3 2019.

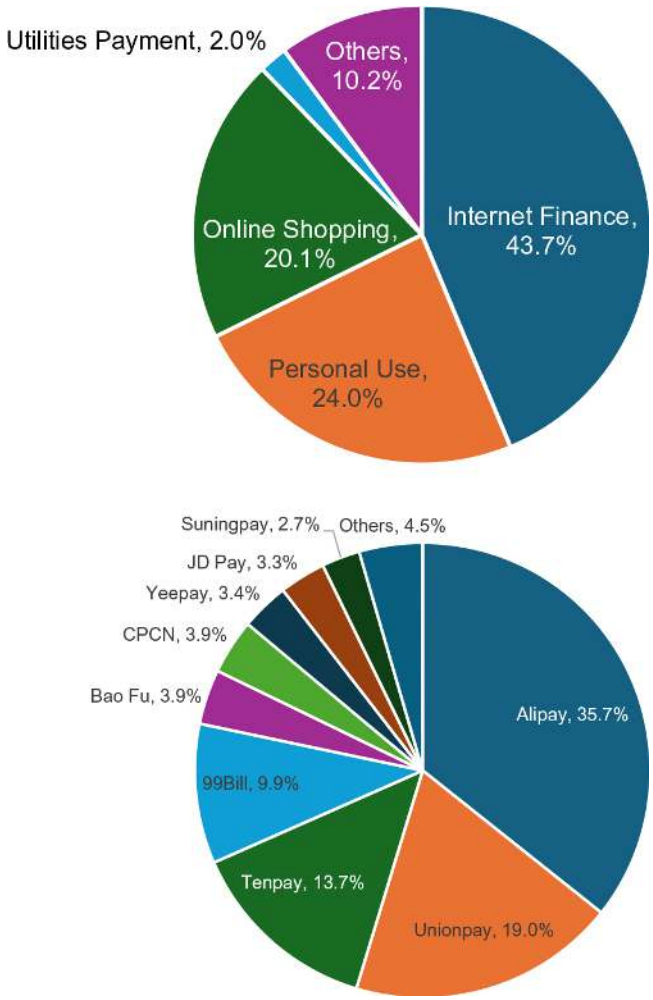


Figure 3.6. China's third-party Internet payment transaction size and market share in Q3 2019.

3.3 Development of P2P Online Lending Platform

Typical P2P online lending involves direct contact between borrowers and lenders on an online lending platform. Investors decide whether to invest by examining the borrower's credit worthiness and loan purpose. The loan service generally includes three main parties: the borrower, the P2P lending service platform, and the investor. The P2P lending platform serves solely as an information intermediary. Borrowers seek funds, and investors seek investment opportunities, which they can directly access on the P2P

platform. This allows for full communication between borrowers and lenders, reducing information asymmetry.

Figure 3.7 shows a sustained growth in the number of new P2P platforms per month from 2016 to 2018, followed by little to no growth in the subsequent years. Figure 3.8 indicates that Beijing, Guangzhou, Shanghai, and Zhejiang accounted for the top four in terms of the number of transactions nationwide, representing 80%–90% of the national volume. Figure 3.9 illustrates the monthly P2P turnover volume from 2016 to 2019, which experienced a steady rise followed by a decline. The comprehensive rate of return also declined from 12% to 10% during this period and stabilized in 2017. Figure 3.10 shows that between January 2016 and September 2017, the number of investors exceeded the number of borrowers. Both groups continued to grow each month, but the growth rate of borrowers was significantly higher than that of investors. However, in September 2017, the situation reversed, and both groups began to decline continuously. In summary, P2P platforms have experienced significant fluctuations, and headwind to its development persists today.

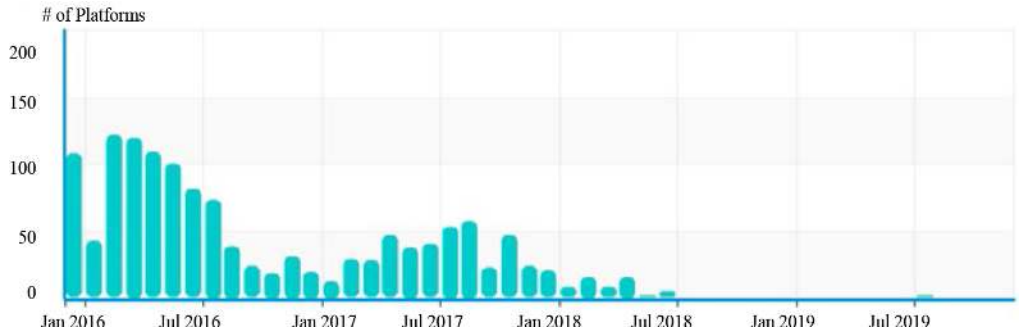


Figure 3.7. Number of new P2P platforms per month from 2016 to 2019.

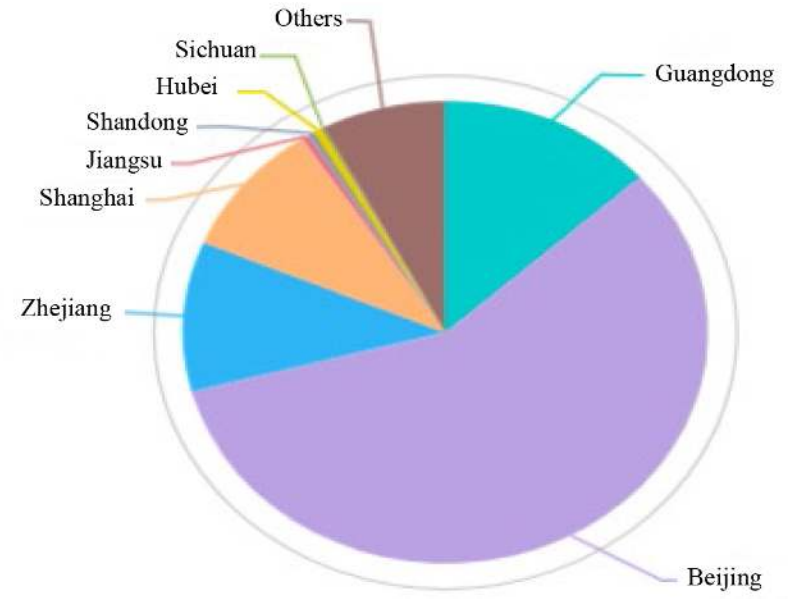


Figure 3.8. Geographic distribution of P2P transaction volume.



Figure 3.9. P2P monthly transaction volume and expected total return rate from 2016 to 2019.



Figure 3.10. Number of monthly P2P investors and borrowers from 2016 to 2019.

3.4 Analysis of the Development of Crowdfunding

Crowdfunding is an Internet-based financial model that originated in the United States in 2009 with the launch of Kickstarter. It involves a collective effort by organizations and individuals who communicate through the Internet to pool funds and support projects. Crowdfunding is typically categorized into four types: debt, equity, rewards, and donations. In China, crowdfunding is divided into four types: public welfare, equity, rights and interests, and property rights. Regardless of the type of crowdfunding, fundraisers set up a personal webpage on the platform's website to introduce the project to investors. After passing identity verification on the crowdfunding platform, they raise funds from the public or seek other types of financial support.

Figure 3.11 illustrates the evolution of China's crowdfunding industry, which has gone through three distinct stages: the nascent stage, characterized by initial growth; the explosive growth stage, marked by rapid expansion; and the industry reshuffle stage, involving adjustments and realignment. Figure 3.12 classifies crowdfunding platforms into several categories: equity-based platforms (such as RenRenChuang and Crowdfunders), reward-based platforms (such as DotFunding and Taobao Crowdfunding), property-based platforms (such as VicC Property Rights and ZhiRenKe), philanthropy-focused platforms (such as Water Drop Fundraising and Tencent LeDonation), and comprehensive platforms (such as Suning Finance and CrowdFunding.com). Figure 3.13 shows the trend in the number of operational crowdfunding platforms. The peak was in 2016, with a total of 532 platforms, with 155 of them being property-type platforms. Afterward, the number of platforms gradually declined across all categories, reaching 105 platforms by the end of June 2019. Figure 3.14 shows that equity-based platforms predominated, accounting for 37%, followed by reward-based platforms at 31%, comprehensive platforms at 13%, property-based platforms at 12%, and philanthropy-focused platforms at 7%. It is worth noting that in the latter half of 2016, there was a surge in automobile crowdfunding platforms, which temporarily elevated property-type platforms to the forefront. However, challenges faced by the industry and closures of platforms have led to a significant reduction in the number.

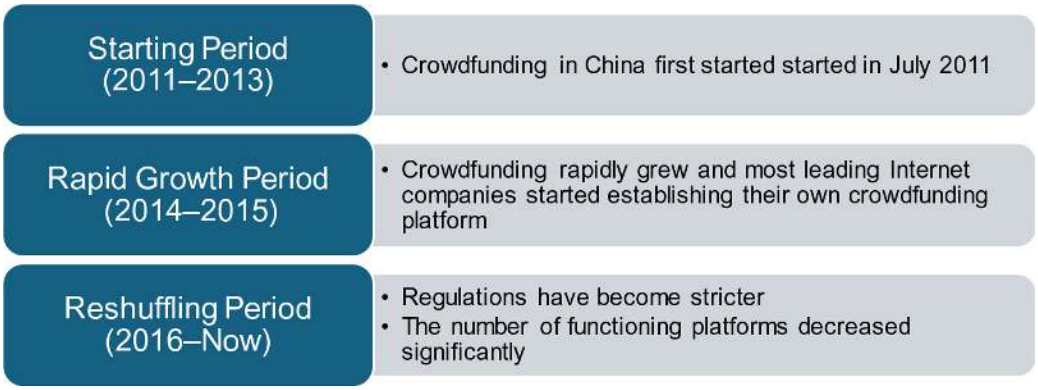


Figure 3.11. The development history of China's crowdfunding industry.

Platform Type	Representative Platform	Characteristics
Equity-based	ZhongChouZhongYuan	The company sells a certain proportion of its shares to ordinary investors. Investors invest in the company to obtain future income.
Reward-based	DotFunding, Taobao Crowdfunding, KaiShiBa, MoDianWang	Reward-based crowdfunding refers to investors investing in projects or companies to obtain products or services.
Property-based	VicC Property Rights, ZhiRenKe, QianCheWang, eZiChan	Raise funds from the public through the Internet to acquire physical assets and obtain profits through asset appreciation and liquidation. The returns can be divided into operating dividends, rental dividends and future value-added income from property rights.
Philanthropy-focused	Water Drop Fundraising, Tencent LeDonation, LvDongWeiLai, MiGongYi	Not for the purpose of profit, but to raise funds from specific people and other public welfare undertakings.
Comprehensive	Duning Finance, CrowdFunding.com, DuoCaiTou	There are many types of projects on these crowdfunding platforms, including projects in smart technology, film and television entertainment, music and books, etc.

Figure 3.12. Crowdfunding platform types.

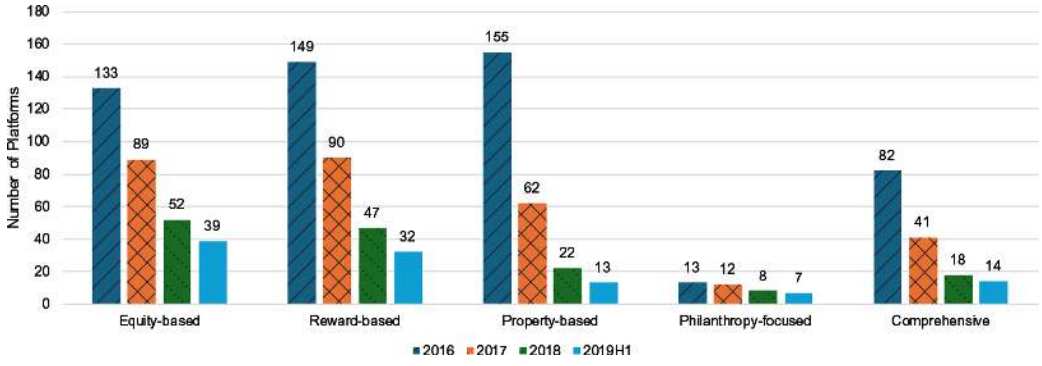


Figure 3.13. Trends in the number of various types of platforms in operation from 2016 to 2019.

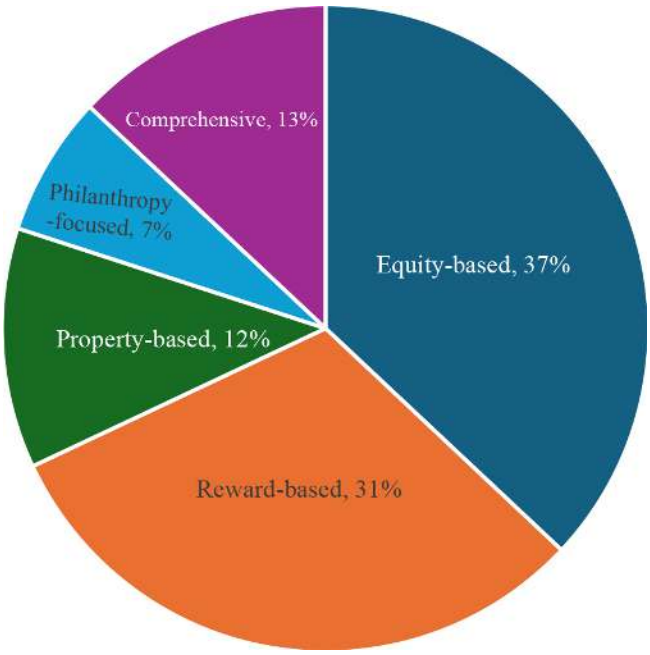


Figure 3.14. Crowdfunding Platforms by Type as of June 2019.

3.5 Current Status of Internet Wealth Management

The integration of the Internet and the financial industry continues to deepen with the rapid development of the Internet. In June 2013, Yu'e Bao was established as China's first Internet money market fund, with Tianhong Fund as the fund manager. Since then, products such as WeChat Wealth Management and Jingdong Xiaojinku have been introduced, and various types of Internet wealth management products have gone online. As a result, the Internet wealth management market has begun to prosper.

Increasingly, investors are no longer satisfied with the rate of return offered by the traditional financial industry and are turning to Internet wealth management. The 2019 China Private Wealth Report, jointly released by China Merchants Bank and Bain & Company, reveals that in 2018, Chinese individuals had investable assets worth 190 trillion yuan. This indicates a vast market space for Internet wealth management.

As of June 2019, the number of Internet wealth management users in China has exceeded 160 million. In 2020, Ant Financial Services, an

affiliate of Alipay, launched a well-known fund investment advisory service called “Help You Invest” in collaboration with the world’s largest public equity fund on Alipay.

Currently, there are three types of platforms on the Internet that offer financial products: traditional financial institutions’ online direct sales platforms, comprehensive online wealth management platforms, and independent professional wealth management platforms. The products offered by these platforms can be categorized into three types based on their level of risk. The first category includes low-risk products such as bank deposits, currency funds, and wealth management products offered by commercial banks. The second category includes medium-risk products such as bond funds and pooled asset management plans offered by brokerage firms. The third category includes high-risk products such as equity, futures, options, and foreign exchange products. With the country’s supervision and interest rate marketization reform, the interest rates of various Internet financial products continue to fall. As of March 2020, the 10-year Treasury bond yield fell by 0.84 percentage points from the previous year’s high. Corporate bond rates fell by about 1 percentage point. The interest rates of some market-oriented deposit products, such as money market funds and structured deposits, have dropped significantly. The money market return rates represented by Yu’e Bao and LiCaiTong have dropped below 2%, which is lower than the effective interest rate of banks’ one-year time deposits. Based on Wind statistics, in 2020, 54% of the bank financial products that disclosed expected annualized returns offered a return rate between 3% and 4%, and 39% provided returns ranging from 4% to 5%. Additionally, 4% of bank financial products featured a yield surpassing 5%. Comparatively, in 2019, the distribution of return rates was 35% for the 3%–4% range, 59% for the 4%–5% range, and 3% for returns exceeding 5%. Therefore, the current yields of Internet financial products will continue to fall, especially low-risk financial products, which no longer have an advantage compared with bank savings interest rates.

Leveraging the Internet as a sales channel for wealth management products has significantly broadened the scope of financial transactions, transforming a substantial number of individuals, often referred to as “long-tail people,” previously overlooked by traditional finance into target investors. This demographic, lacking adequate investor education, exhibits

limited proficiency in risk identification and low-risk tolerance, rendering them more susceptible to individual and collective irrational behaviors.

3.6 Internet Financial Risks

3.6.1 *Network security risks*

Contemporary society is deeply entrenched in advanced electronic information technology, with the Internet serving as a convenient platform. However, this technological landscape also harbors numerous potential security risks, and the exploitation of these risks by malicious entities can lead to unforeseeable losses. The primary areas of risk lie in third-party software, protocols, and network vulnerabilities.

When engaging with payment platforms such as Alipay and WeChat, users unknowingly expose their personal information to third-party organizations. In the event that malicious actors employ phishing websites or false information for transactions, the divulgence of personal privacy can result in incalculable losses, leading to financial risks. Furthermore, conducting fund transfers through bank apps in public networks heightens the probability of fraud. If connected to a public network established by criminals, this provides an opportunity for them to exploit Trojans or viruses, posing a continuous threat to users' devices and, consequently, amplifying financial risks. Thus, the pervasive use of electronic payment platforms and bank apps necessitates heightened vigilance to mitigate the evolving challenges of network security risks.

3.6.2 *Information asymmetry risk*

The risk of transactional information asymmetry is heightened in Internet wealth management. Third-party wealth management applications, such as Alipay and WeChat, offer various products with a seven-day return rates ranging from about 2% to as high as 7–8% or even higher. However, it is widely known that high returns correspond to high risks. Although these packaged products can bring super high returns, they also come with high risks. The asset allocation composition of these products is often unknown. It is unclear whether there is any restructuring of non-performing assets

behind these products with extremely high returns. Similar to the subprime mortgage crisis, various subprime mortgages or junk bonds were restructured into bonds with better credit ratings and then resold, which brought unpredictable consequences to investors. Therefore, information asymmetry and opacity can bring financial risks.

3.6.3 *Default risk*

Internet finance offers a virtual credit lending platform that enables customers to secure small loans promptly to address immediate financial requirements. This is predominantly facilitated through banking institutions as intermediaries. Additionally, third-party institutions may assume responsibility for guaranteeing credit lending activities. The key concern in the realm of credit lending within Internet finance revolves around the risk of default. Given the inherent instability and virtual nature of Internet finance, potential instances such as bankruptcy, liquidation, or misappropriation of funds by capital-raising institutions can inflict substantial harm upon investor interests. Moreover, these events have the potential to introduce a notable degree of volatility to the financial industry, thereby negatively impacting the continued advancement of Internet finance.

3.6.4 *Risks of imperfect regulatory system*

Every industry, including the financial sector, operates within the bounds of legal frameworks, and the Internet finance industry is no exception. Nevertheless, Internet finance diverges from traditional finance due to its rapid development over a relatively short period, resulting in a regulatory framework that is still evolving and exhibiting limitations in its coverage and application. While traditional finance benefits from a well-established regulatory system safeguarding the interests of transaction parties, the burgeoning nature of Internet finance creates vulnerabilities and shortcomings in its legal coverage. While thousands of people enjoy the benefits of Internet finance every day, there are also many criminals waiting for opportunities behind the scenes. The healthy development of the financial industry may face significant risks if Internet finance lacks a

sound legal system as a protective umbrella. This could eventually lead to a decline in the industry.

3.7 Conclusion

China's Internet finance sector, which includes third-party payments, peer-to-peer lending, and crowdfunding, faces significant obstacles that hinder its growth. Overcoming these challenges is crucial to unlocking the industry's full potential. Research aimed at surmounting these barriers is essential to fostering expansion. Financial technology (fintech) is a pivotal tool in this pursuit. The use of advanced technologies such as blockchain, big data, and artificial intelligence (AI) shows the potential in overcoming obstacles and driving the industry toward greater success.

Chapter 4

Application of Machine Learning in Internet Financial Risk Research

4.1 Current Status of Internet Financial Regulation

While the potential benefits of Internet finance are evident, it is imperative to acknowledge the associated risks it poses to the financial system. The expansion of traditional financial services onto the Internet, along with the emergence of Internet-based financial services, has given rise to new financial institutions. However, the absence of regulatory frameworks has created a regulatory void, particularly concerning Internet intermediary service models. This regulatory ambiguity presents a significant risk within the Internet finance sector.

4.1.1 *External regulation and industry self-regulation*

China's regulatory system lacks specific provisions for Internet financial businesses, resulting in a regulatory gap. Prompted by incidents like the collapse of "Haha Loan" and emerging industry risks, the China Banking Regulatory Commission (CBRC) issued the "Notice on Risks Related to Renren Dai" in August 2011. This document, while acknowledging that P2P credit operations fall outside the CBRC's purview, urged commercial banks to establish "firewalls" to shield against potential transmission of P2P credit risks to the banking system. However, industry self-regulation remains lax, with few institutions adhering to the Self-Discipline Convention for Individual-to-Individual Microcredit Information

Consulting Service Institutions issued by the China Microfinance Union in January 2013.

4.1.2 *Credit information exchange*

In countries like the United Kingdom and the United States, where P2P lending originated, borrower registration typically involves providing extensive personal and financial information, leading to robust credit information sharing systems. In contrast, China's credit environment lacks integration with Internet finance companies, hindering effective credit information exchange. Consequently, Internet finance firms rely heavily on internal credit assessment mechanisms, resulting in inadequate *post-facto* disciplinary measures and low borrower default costs, posing significant risks to the industry's long-term viability. Furthermore, the absence of credit information exchange can result in significant delays and inefficiencies in Internet finance companies' efforts to independently gather customers' credit and financial data. This delay may exacerbate issues such as malicious lending practices and the cycle of repaying old loans with new ones, posing substantial risks within the industry.

4.1.3 *Technology and platform security*

Not only has the integration of financial services with Internet technology undoubtedly enhanced business convenience, but it has also amplified information and capital security concerns. Despite advancements, both traditional and Internet financial platforms face technical vulnerabilities. Internet financial firms, with relatively limited resources compared to traditional institutions, heavily rely on their network platforms for credit assessment and risk management, exacerbating platform vulnerabilities.

With rapid technological advancements, artificial intelligence (AI) and big data are increasingly prominent in various industries, particularly in finance. The development of big data in Internet finance offers new avenues for credit risk mitigation in an era where traditional data techniques fall short.

4.2 Internet Financial Risk Control in the Context of Big Data

4.2.1 *Emergence and characterization of big data technologies*

The Internet, as a platform for information dissemination and regeneration, generates a wide variety of data types and formats, including not only structured data but also semi-structured and non-institutionalized data. The scale of multi-source heterogeneous data is huge, and the use of massive social credit data has gradually penetrated business operations, bringing new opportunities for risk information identification in financial enterprises. In the financial industry, the reliability of risk prediction can be improved by analyzing a large amount of data in multiple fields in a non-linear way. Problems such as information security and information processing that arise with big data technology also add to the difficulties of business management, making it difficult to make quick and correct judgments. Enterprises need to use big data technology to distill valuable information and establish risk models based on big data in order to better utilize big data technology empowerment to improve the accuracy of Internet financial risk control.

4.2.2 *The significance of the application of big data technology in Internet financial risk control*

Big data technology is a powerful tool to promote the construction of the Internet financial risk control system. The development of Internet financial risk control technology started late, but the sophistication of technology continues to progress, and the use of artificial intelligence is gradually unfolding. Internet risk control technology is subject to system barriers and the uncertainty of personal credit, the network of a large amount of social information collection and data program processing also requires long-term tracking and monitoring, so to achieve the intelligent development of the network, Internet financial risk control needs to be continuously improved. Currently, the Internet financial industry, in the era of the transition is faced with a large number of data value mining tasks. With the development of the information age, computer technology, and science and technology level, the processing of big data technology has been gradually perfected;

technologies such as data mining, extraction and processing are not limited to following protocols but have also realized data statistics and risk control, laying the foundation for Internet financial risk control. Big data has been used to empower authentication tools such as personal identifiable information screening and facial recognition. These tools when coupled with guarantees provided by financial institutions can make risk control in Internet finance more effective.

Taking credit risk and fraud risk, to which the Internet finance industry attaches more importance, as an example, in order to form a basic model of Internet finance risk control based on big data, it is necessary to verify and optimize it by continuously acquiring a large amount of new data. The basic idea is to compare the borrower's past credit history information with the credit habits of all borrowers in the database and check whether the borrower's development trend is similar to that of various borrowers in financial difficulties who often default, overdraw arbitrarily, or even file for bankruptcy. Many domestic banks have begun to try to drive business operations through big data, such as CITIC Bank credit card center using big data technology to achieve the actual marketing; China Everbright Bank has established a social network information database, and China Merchants Bank is using big data to underwrite small and microloans.

4.2.3 *Strategies for applying big data technology in the field of Internet financial risk control*

The utilization of big data underscores the pivotal role of data in the Internet era, particularly as attention increasingly focuses on the Internet credit system. Therefore, financial institutions must refine their big data risk control strategy.

Database Construction: Strengthening the foundation of Internet financial risk management requires a concerted effort toward constructing robust databases and integrating database resources. Big data is not just about arithmetic laws and machines. Humans still play an important role in judgment, screening, knowledge learning, engineering and technology creativity. This human element is essential for realizing diverse scenarios in Internet finance through data simulation, thereby expanding the field's potential and facilitating effective financial risk control.

Continuous Big Data Algorithm Optimization: While big data serves as both a resource and a technical tool, it demands ongoing refinement for optimal utilization. Financial institutions must prioritize the enhancement of big data algorithms to accurately and efficiently sift through vast databases, extracting pertinent information swiftly. This necessitates advanced computational techniques and methodological precision, fostering the integration and optimization of resources to propel the evolution of Internet financial risk control.

Development of External Sensing Technology: To maximize the efficacy of big data in Internet financial risk control, efforts should extend beyond database integration and algorithmic enhancement. Strengthening the development of financial risk sensing technology is imperative. This entails bolstering the comprehensiveness of financial risk management, aligning with market demands, and establishing clear objectives to underpin research on the application of big data in Internet financial risk control.

4.3 Internet Financial Risk Control and Machine Learning

The credit business is the primary source of profit for commercial banks and Internet finance companies. Internet finance is the extension of traditional credit business in the online scenario. For users with a need to borrow, Internet platforms need to evaluate their creditworthiness and ability to repay and lend to users with the highest repayment probability to maximize profits. The primary objective of Internet finance is risk control based on credit business. In the credit field, there are two main types of risks: credit risk and fraud risk.

Credit risk refers to the possibility that a borrower may become unable or unwilling to repay a loan after it has been taken out. This risk is often associated with changes in the borrower's financial situation or state of mind due to unforeseen circumstances.

Fraud risk, on the other hand, arises from the borrower's improper use of the loan, resulting in a lack of repayment plan from the outset.

Compared to fraud risk, credit risk is more controllable because financial institutions can prevent credit risk through risk pricing strategies and other means. Fraud identification is crucial for risk control. Purposeful borrower fraud can result in significant losses for financial institutions, especially in cases of group fraud. Identifying and preventing fraud is essential to avoid severe damage to the credit business. It is important to maintain a clear and logical structure when discussing lending risk control.

Financial institutions rely on credit scoring systems and fraud detection systems to control credit and fraud risks. Both systems are crucial in managing risk in the credit field. Credit scoring systems assess the borrower's ability and willingness to repay, while fraud detection systems evaluate the borrower's purpose of borrowing. Financial institutions face tens of thousands of loan requests and large amounts of user borrowing and lending data daily. Traditional manual credit reviews cannot meet these demands. However, an AI risk control model based on machine learning can quickly and accurately process loan applications in batches. It can also output a person's credit score and fraud score, freeing up the labor force of credit reviewers in small and medium-sized loans.

Machine learning-based Internet financial risk control is one of the primary applications of AI at present. The AI models commonly used in Internet financial risk control systems include fraud detection models, approval models, credit models, and risk pricing models. Among them, the fraud detection, borrower approval, and credit models belong to the binary classification model. The commonly used modeling method is the supervised learning model of machine learning. Additionally, the fraud detection model can also be used to detect group fraud with the help of social relationships among loan users. Risk pricing models contain some linear programming-related content, but its relationship with machine learning is weaker, so this will not be the focus of this chapter.

4.4 Application of Machine Learning in Credit Scoring

As data mining technology advances, the integration of machine learning techniques for risk management has become standard practice in the Internet finance domain. Within Internet finance, the two primary risk management frameworks which heavily leverage machine learning are credit scoring and fraud detection. Credit scoring aims to construct a credit scorecard, delineating the creditworthiness of individual investors, while fraud detection primarily involves categorizing fraudulent activities perpetrated by individual borrowers.

4.4.1 *Theoretical model of credit scoring*

The main purpose of the credit scoring model is to measure a customer's credit risk, i.e., the ability and willingness to repay. Compared with the traditional logistic regression scoring model, the scorecard based on the ensemble methods in machine learning performs better in key indicators, such as KS value and capture rate. Take the integrated model scorecard based on XGBoost as an example. Since XGBoost is not as sensitive to input as the logistic regression model, feature processing will be simplified, but its parameter adjustment process is more complicated. When measuring user credit risk in the field of Internet financial risk control, we will start from the factors shown in Table 4.1.

Table 4.1. Factors affecting credit risk.

Factor	Example	Risk
Demographic information	Age, gender, family situation, etc.	Being too young may signify a lack of income source. Being too old may pose a lifespan risk. Women often exhibit a stronger inclination toward repayment.
Financial information	Income, household assets, etc.	Data that directly reflects the user's repayment ability.
Loan information	Users' loan status across multiple borrowing platforms	Exceeding borrowing thresholds may indicate poor repayment capability.
Consumer behavior	E-commerce data, travel data,	Reflects consumption levels and purchasing power. Elevated spending may warrant scrutiny.

	takeout data, etc.	
Historical platform performance	Repayment history on borrowing platforms	Calculate the sum of the user's historical maximum and minimum overdue days and historical borrowing amount.
External credit data	Credit information on the market	Extract credit scores to measure user credit risk.

As a measurement tool, the credit scoring model needs to be calibrated like weights, i.e., giving each unit an appropriate scale. Therefore, after model training, it is usually necessary to use a unified scoring mapping to normalize the model's output. Lending institutions generally have the following requirements for scoring tools:

(1) Attribute points are positive.
(2) Total scores are positive.
(3) Each score represents a specific credit quality.
(4) Score differences represent uniform changes in credit quality.

In order for the scorecard model using machine learning to meet the above conditions, proportional scaling score mapping can be employed. Initially, assume a user's base score to be 600 points. Subsequently, when the user's non-overdue probability is twice the overdue probability, 50 points are added. Similarly, if the non-overdue probability is four times the overdue probability, 100 points are added. This pattern continues, with the non-overdue probability exceeding the overdue probability by eight times resulting in the addition of 150 points. Following this methodology, we derive the industry standard scorecard conversion formula:

$$\text{Score} = 600 + 50\log_2\left(\frac{P_{\text{positive sample}}}{P_{\text{negative sample}}}\right) \tag{4.1}$$

where Score is the output after scorecard mapping, $P_{\text{positive sample}}$ is the probability that the sample is not overdue, and $P_{\text{negative sample}}$ is the probability that the sample is overdue.

Another mapping method does not consider the true meaning of score conversion, such as the credit score model based on XGBoost. In this model, the output probability value represents the likelihood of the user

being overdue. Consequently, as the score approaches 1, the probability of overdue increases. Therefore, the XGBoost score mapping formula is straightforward and direct:

$$\text{Score} = 600 + 50\log_2\left(\frac{1-\text{pred}}{\text{pred}} - \text{lag}\right) \tag{4.2}$$

where pred is the likelihood generated by the XGBoost integrated model and lag is usually set as the corresponding probability when the predicted score is equal to the base score. Assuming that the 20% quantile of the model's predicted scores defines the base score, the value of lag will equal the predicted probability corresponding to the sample which ranks in the 20% quantile on the test set.

The XGBoost model supports custom evaluation functions and loss functions. As long as the loss function is second-order differentiable, the model parameters can be solved by maximizing the evaluation function. It can be adjusted according to specific business goals in real-life scenarios.

4.4.2 *Research on credit scoring model construction*

As a major form of Internet finance, P2P platforms have many legal loopholes, and thus, enhanced regulatory measures are necessary. Certain platforms capitalize on investors' desire for high returns and expand rapidly, but at the same time, they are accompanied by exceptionally high default risks. To foster the sound development of Internet finance, platforms such as P2P must establish effective and comprehensive credit scoring models. With the advent of the big data era, risk control strategies leveraging technologies like machine learning have emerged, with a focus on proficiently processing platform data and transforming vast amounts of unstructured and semi-structured data into standardized formats to optimize data utilization.

In 2011, Professor Bachmann and others conducted research on the development and application of big data technology in P2P online lending platforms.[18] In the same year, Manyika and others from McKinsey & Company conducted research and analysis on the development and application of big data in the finance industry.[19]

Jiang Mingzhe (2015) believes that big data technology holds immense application potential in commercial banks, potentially replacing the traditional credit approval processes with customized risk mitigation methods tailored to each credit scenario. This necessitates an accurate analysis of individual credit positions using big data technology.[20] Lin Yin *et al*. (2015) delved further into bank bad debt issues, advocating for the use of big data technology to analyze historical non-performing loan data, thereby establishing a robust risk management system by combining the APH algorithm with the application of BSC theory in risk control.[21]

Regarding risk control algorithms for P2P online lending platforms, Malekipirbazari *et al*. (2015) conducted extensive research on borrower credit assessment methods, revealing the efficacy of random forest-based techniques in identifying low-risk borrower groups.[4] However, this method's drawback lies in occasionally misclassifying sub-optimal customers as high-risk borrowers during credit ranking. Guo *et al*. (2016) proposed using kernel regression for credit risk assessment in P2P lending, enabling not only individual loan risk quantification but also borrower risk estimation based on past lending behavior.[6]

Jin *et al*. (2015) first used the random forest model and correlation coefficient matrix to identify the 10 most important credit evaluation indicators, categorizing prediction results into default, on the verge of default, and paid off (not default) groups. This diverges from conventional risk prediction models by adopting a nuanced approach. The paper inputs the borrower's relevant basic information, such as credit risk and income level, into various models and compares the predicted results of the random forest model with those of other models.[5]

Chen Xiao's research (2014) on borrowers' default probability and motivation post-loan acquisition revealed tendencies to falsify essential information such as their financial conditions and credit score to secure loans.[22] Xia *et al*. (2015) believes that there are a large number of non-professional and irrational investors in P2P online loans, and their participation increases the default rate of online loans, which results in a large amount of data noise in the obtained borrower risk information.[15] Therefore, Xia and others researched and introduced a new model to reduce the interference of data noise on model performance, reduce the impact of subjective selection of the initial clustering core and number of clusters on

the model accuracy in the clustering stage, and introduced the cost of misjudgment as a metric to evaluate the model performance.[11]

Ding Lan and others (2017) studied the default risk alert system of P2P online lending platforms, using logistic regression, decision tree,[23–26] and support vector machine (SVM)[27,28] as primary learners to build an integration strategy based on Stacking to predict the default risk of borrowers in P2P online lending.[29, 30]

In the traditional loan model of financial institutions, borrowers often need to provide collateral to obtain credit lines, and big data uses its own technological advantages to completely disrupt the traditional credit transaction model and largely solve the information asymmetry between lenders and borrowers. This transformation not only enhances financial efficiency and reduces costs but also facilitates cross-industry integration, thereby reshaping the landscape of the entire financial market.

4.5 Application of Machine Learning in Fraud Detection Systems

Internet financial risk control mainly relies on two technical means: credit scoring and fraud detection. Within this risk control framework, fraud detection systems leveraging machine learning technology stand out as pivotal tools widely adopted in the Internet finance domain. Fraud detection includes both individual fraud and group fraud, each presenting distinct challenges. Individual fraud, characterized by a very small proportion of the population that is significantly different from the population, is similar to outliers. Consequently, anomaly detection techniques are often deployed for individual fraud detection, transforming the task into a classification problem within an imbalanced dataset. Practical success in this area often requires relevant business experience. On the other hand, group fraud poses significant threats, causing substantial losses and disruption to Internet financial institutions. Community discovery algorithms need to be combined with machine learning methods to effectively detect potential risks.

4.5.1 *Theories related to individual fraud*

Since fraud samples account for a very small proportion in real scenarios, and the ratio of positive and negative samples can reach 100,000:1, the training effect of supervised learning cannot meet the requirements. Even if new samples are synthesized through oversampling technology such as the SMOTE algorithm, problems still exist. Since fraud methods develop and change over time, the identification effect of models obtained through historical fraud samples is not obvious when dealing with new fraud methods. Therefore, other models need to be used to learn the special properties of fraud samples.

To solve the problem of individual fraud detection, a better solution is to regard fraud samples as outliers. The outliers are obviously different from most of the observed values and are part of the sample set. They may be caused by real data and may also be generated by noise. Noise is defined as the random error that caused the data to deviate from the true signal or underying pattern. It can be regarded as an outlier, but it may not meet the standard of outliers. Outliers are objects in the dataset that deviate from the trend of the overall dataset. Usually, the degree of anomaly of outliers is greater than that of noise. However, in practical applications, it is difficult to distinguish between the two.

4.5.1.1 *LOF anomaly detection algorithm*

Local outlier factor (LOF) is a density-based anomaly detection method. The algorithm is simple and intuitive. Compared with traditional anomaly detection methods, this algorithm does not require the assumption that the data obey a specific distribution (since the assumed distribution is usually not true in reality) and can also output anomaly scores well.

The LOF algorithm quantifies the distance between each sample and its k nearest neighbor samples into a concept of density, which is called local reachable density. First, the number k of nearest neighbor samples needs to be determined, and then the anomaly score is obtained according to the formula. Considering that there may be duplication between samples and the distance between duplicate samples is 0, the sample space will be deduplicated at the beginning of the algorithm.

The algorithm flow is as follows:

(1) First, the sample space is deduplicated, and the distance between each sample and the remaining points in the sample space is calculated separately.
(2) Arrange the distances in step 1 in ascending order.
(3) Specify the number of neighbor samples *k*, and for each sample point, find its *k* nearest neighbor samples, and then calculate the LOF score, which is the anomaly score. This anomaly score is calculated based on the local reachable density.

4.5.1.2 *IF anomaly detection method*

Isolation Forest (IF) is an algorithm based on the idea of random partitioning of space. It is obtained in parallel from multiple binary trees and then the output results are weighted and averaged. In the traditional binary tree, features and corresponding split values are chosen at each level to minimize the standard error and this process is applied recursively to obtain the final decision tree. In each isolated tree of IF, the features and feature values are completely randomly selected from the data.

The algorithm flow is as follows:

(1) Randomly select a certain number of the samples from the sample space and randomly select a feature from the feature space, i.e., row–column sampling.
(2) Randomly select a feature value from the existing feature dimension as the dividing node, i.e., the threshold.
(3) "Split the decision tree" so that the left branch is placed with samples less than or equal to the threshold, and the right branch is placed with samples greater than the threshold.
(4) Repeat the above process until the data are no longer separable (e.g., all samples of the current leaf node have the same value in all feature dimensions) or the differentiation of the current tree reaches the binary tree depth set at the beginning.

4.5.2 *Related theories of group fraud*

In finance, aggregation means risk. The central idea of group fraud detection is group discovery. Knowledge graph is the main means to identify group fraud. It uses a graph-based data structure to store knowledge in a graph and returns processed and inferred associated information.

4.5.2.1 *Basic concepts*

For group fraud detection, the most effective method in the industry is the community discovery algorithm based on graph structure, which is mainly divided into two categories:

(1) **Community detection:** By partitioning nodes in a fixed network into smaller groups, we obtain clusters that represent communities.
(2) **Community search:** Given an online community and a certain node A in it, start from node A to find communities related to A.

For a community in a network, the points within the community should be closely connected, while the points between communities should be quite different. It can be seen from this characteristic that the community discovery algorithms are similar to clustering, and they also belong to the category of unsupervised learning. The most important measure in clustering is the similarity between samples, and this is also true in communities. Commonly used similarity measures include Jaccard similarity and Cosine similarity.

After defining the characteristics of each sample and the indicators for calculating sample similarity, the similarity between each sample can be calculated separately, thereby dividing users into different communities through thresholds. It is worth mentioning that the traditional clustering algorithm for community division is not stable and has poor accuracy in most cases.

4.5.2.2 *Girvan–Newman algorithm*

The Girvan–Newman (GN) algorithm is one of the most classic community discovery algorithms. The basic idea is to remove the edges connecting communities, and the remaining structure is the community. The shortest path between points in two communities always passes through the connecting edges between communities. Therefore, the edges with greater centrality are more likely to be connecting the edges of different communities. Therefore, you only need to continuously remove the edges with large centrality to get independent communities.

The algorithm steps are as follows:

(1) Calculate the betweenness centrality of all edges.
(2) Remove edges with the highest betweenness centrality.
(3) Recalculate the betweenness centrality of the edges affected by the removed edges.
(4) Repeat steps 2 and 3 until the number of disconnected communities reaches the preset value.

According to this idea, as the edges gradually decrease, the points in the connected graph can certainly be divided into many disconnected sets.

4.5.2.3 *Louvain algorithm*

Although the GN algorithm has high accuracy, it has high computational complexity. The Louvain algorithm is more commonly used for community discovery. Before introducing this algorithm, you need to understand the concept of modularity. For a given connected graph, there can be many division methods. To determine which division method is the best, you need to use the concept of modularity.

First, define

$$A_{ij} = \begin{cases} 1, & i \text{ and } j \text{ are connected} \\ 0, & i \text{ and } j \text{ are not connected} \end{cases} \tag{4.3}$$

$$\delta(c_i, c_j) = \begin{cases} 1, & i \text{ and } j \text{ are in the same community} \\ 0, & i \text{ and } j \text{ are not in the same community.} \end{cases} \tag{4.4}$$

Redefine modularity Q:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{4.5}$$

.

Among them, m represents the number of edges in the network, k_i and k_j represent the degree of i and j, so $\frac{k_i k_j}{2m}$ is the expected value of the sum of the numbers of i and j. The larger the modularity Q, the better the result of community division is. When modularity $Q > 0$, the current divided module is better than a random division. The industry generally believes that when $Q > 0.3$, the current division is a better division.

The Louvain algorithm can be understood as the inverse process of GN: The idea of GN is to continuously remove edges, similar to top-down hierarchical clustering, while the Louvain algorithm is constantly condensing, similar to bottom-up hierarchical clustering.

The algorithm steps are as follows:

(1) Treat each node as an independent community.
(2) If adding i and j to the current community will increase their modularity, compare the two values and choose the one with the larger increment to join the current community.
(3) Iterate repeatedly until the value of modularity Q no longer increases.

4.5.3 *Related research on fraud detection models*

With the rapid development of "Internet + Finance," fraud detection has always been the focus of financial institutions and an important research topic for domestic and foreign researchers. Among the issues, the problem of sample imbalance is a key factor affecting the accuracy of fraud detection, and the detection accuracy of existing fraud detection models can be further optimized and improved.

Research on the identification and management of credit defaults started early outside China, mainly focusing on the determination of default behavior based on predetermined rule bases and traditional machine

learning algorithms. In 1980, Wiginton *et al*. compared the performance of consumer credit scoring models based on linear discriminant analysis and logistic regression, and the results showed that the latter had higher predictive accuracy.[31] In 1992, Nath *et al*. used commonly used mathematical programming methods and regression analysis to conduct credit modeling, and experiments showed that the performance of the two algorithms was relatively close.[32] In the same year, Davis *et al*. compared and analyzed the top-down inductive learning algorithm with the multi-layer perceptron and backpropagation neural network algorithms, and they were verified on the Bank of Scotland credit dataset.[33] Since the rise of data mining research in the 1980s, artificial neural networks and support vector machines have been practically used in the field of credit default identification.[34] In 2011, Bhattacharyya *et al*. comparatively analyzed the logistic regression algorithm, support vector machine algorithm, and random forest algorithm in the identification of fraudulent credit card transactions. The results showed that the logistic regression algorithm was inferior to the other two algorithms in terms of specificity and accuracy.[35] There are still many studies in the financial field based on ensemble learning.[36] Since data in the financial field often exhibit unbalanced characteristics, anomaly detection methods are also commonly used. Among them, Yamanishi *et al*. (2004) targeted the problem of medical insurance fraud by using the unsupervised SmartSifter algorithm and Hellinger distance to predict fraud behavior.[37]

While research on credit default identification in China mainly focuses on the construction of credit risk scoring models, there is a lack of research on the prevention and control of fraudulent transactions. In 2006, Yan Hua, Hu Mengliang, and others conducted data mining on historical credit card transactions and constructed a credit card anti-fraud model based on the Bayesian method to identify whether credit card customers have experienced transaction fraud.[38] In 2008, Yang Xi built a fraud risk detection model based on support vector machines in order to improve the recognition rate of abnormal transactions. Guo Tao and others proposed to first calculate the confidence level of the training data and then use the processed data as new samples to train the neural network model for better feature learning and then introduce the ROC analysis method to formulate thresholds to identify outliers.[39] In order to improve the fraud detection rate

and effectiveness, the proposed credit card anomaly detection system has been experimentally verified and has good practicability and adaptability.[40]

In terms of fraud detection algorithms, some Chinese scholars have also successively proposed some improved algorithm models. In 2013, Yang Yi realized credit card fraud detection by combining clustering and AdaBoost algorithms.[41] Chen Min (2016) proposed a fraud detection method based on fuzzy LOFs.[42] Su Yating (2017) proposed credit card fraud detection based on the fuzzy two-norm kernelless quadratic support vector machine.[43] In response to the extremely imbalanced problem of credit card fraud detection datasets, Anis (2018) proposed to use neighbor resampling and classification sorting to process imbalanced datasets, thereby improving the performance of the fraud identification model.[44]

By analyzing and studying the current research, we found that Internet financial fraud identification based on machine learning technology has become a current research hotspot. However, the existing identification technology still has some shortcomings, which are mainly reflected in the following aspects:

(1) There is a serious shortage of labeled samples. Supervised credit default identification methods are weak in identifying new types of default forms. In a big data environment, the cost of labeling massive data is high, and credit default identification models need to be efficient and timely in order to effectively detect fraud.
(2) The fraud detection rate needs to be improved. With the rapid development of Internet finance and big data, it has become increasingly difficult to extract fraud characteristics from massive data, and the diversity and variability of fraud have made it difficult to identify credit defaults; secondly, extreme imbalance of data often leads to low classification accuracy. The imbalance problem becomes particularly important in default identification. Accurate prediction of minority class defaults can avoid bad debt losses to a large extent.
(3) Lack of adaptability, which means the ability to identify and predict unknown default behaviors, is a key issue that needs to solved. Existing technologies for modeling credit defaults still exhibit limitations in this regard and warrant further research and enhancement. Therefore, we need to build independent feature learning capabilities and ensure good

adaptive capabilities. The credit default identification model is also the focus of this topic.

(4) The efficiency of processing credit big data needs to be improved. In the big data era, credit default identification requires identifying fraudulent transactions from large-scale data. Accelerating the detection of fraudulent transactions while ensuring a high detection rate is a major problem that needs to be solved.

(5) The data type is complex. Online transactions usually occur on websites and smart mobile devices. During the data preprocessing stage, data from different devices need to be normalized so that it can be evaluated fairly across platforms. Additionally, the datasets collected from e-commerce usually have high dimensions, which further complicates the data processing task.

Deep learning can overcome the above shortcomings due to its excellent feature extraction and powerful adaptive capabilities. Therefore, the application of deep learning technology to credit default identification has gradually become favored by a large number of scholars and has become one of the current prominent research directions. The following chapter will introduce in detail the application of deep learning technology in Internet risk control.

Chapter 5

Application of Deep Learning in Internet Financial Risk Research

5.1 Background Introduction

Internet finance is an organic combination of financial functions and Internet technology. Through information and communications technologies, such as big data, cloud computing, social networks and search engines, a new financial business including financing, payment, investment and information intermediary services, is realized. In this scenario, market information transparency can greatly reduce the disadvantages caused by information asymmetry, reduce intermediate processes, reduce transaction costs on both the supply side and demand side of funds, and improve the efficiency of capital operation.

The development of Internet finance in China has roughly gone through three stages:

(1) **Stage of Internet-Enabled Finance:** The combination of finance and the Internet at this stage is mainly reflected in the use of Internet technology by traditional financial institutions to move offline businesses online to realize the automation and electronicization of financial services, such as banks' core transaction systems, credit systems, clearing systems, etc. At this time, the Internet financial industry has not yet formed in the true sense.

(2) **Stage of Substantive development of Internet finance:** With the rapid development of third-party payment, the integration of the Internet and finance has deepened from the technical field to the

financial business field. The most representative ones are P2P online lending represented by Renrendai and CreditEase, JD Crowdfunding, Crowdfunding platforms represented by Zhongchou.com and Angel Investment, as well as comprehensive Internet financial service platforms led by Ant Financial, JD Finance, Lufax, and Baidu Finance. At this stage, the asset side, transaction side, payment side, service side, and capital side of the financial industry have all realized interconnection, information sharing, and disintermediation through the Internet and mobile terminals.

(3) **Stage of Technology-Driven Finance:** Technology continues to drive transformative changes in the financial industry. Advanced IT technologies such as big data, cloud computing, artificial intelligence, and blockchain, enable richer and more accurate information collection, the development of more personalized and targeted risk pricing models, a more rigorous investment decision-making process, and a more transparent and fair role of credit intermediaries. These technological advancements greatly improve financial business efficiency and quality of services. Representative applications include big data credit reporting, robo-advisory, and supply chain finance. At this stage, financial technology emerged as a key driver of Internet finance, raising important questions about how big data and artificial intelligence will interact with finance and the potential future changes they may bring to the industry.

5.2 Application of Artificial Intelligence in the Field of Internet Finance

What is the mechanism of applying artificial intelligence to Internet finance? First of all, big data technology is used to collect, process and analyze massive data, and each user is given a characteristic value based on the analysis results, which are then used to establish user behavior models. The results are then calibrated by repeatedly training the model on large amounts of data. Finally, the model can be used to make corresponding reasonable predictions for new individuals. In the following sections, we will elaborate on specific applications of AI in several aspects.

5.2.1 *Risk control*

The core issue in the financial industry is risk control. There are risks in both lending and investing. The higher the risk, the greater the return. Finding a balance between risk and return is crucial. Take credit as an example. Internet credit is usually small and scattered. It is impossible to conduct a door-to-door credit investigation on such a group of users. Moreover, credit data obtained through traditional methods are relatively limited and typically include only a small set of characteristics, such as customer's occupation, income, age, and fixed assets. In contrast, artificial intelligence can analyze massive datasets to mine user information and construct high-dimensional user data with thousands of features, thus completing a user portrait of a lender and offering personalized risk control suggestions. Therefore, the application of artificial intelligence represents a qualitative leap forward in the accuracy and efficiency of Internet financial risk control.

5.2.2 *Anti-fraud*

With the development of society, various fraud methods have emerged one after another and are difficult to guard against. Individual fraud has rapidly evolved into organized and large-scale group fraud and related risks. Risk control in the financial industry is facing severe challenges. In this case, artificial intelligence's machine learning and knowledge graph will come in handy. For example, artificial intelligence can process large-scale and high-frequency data to construct various relevant information nodes of the applicant into a huge network graph and, on this basis, establish an anti-fraud model based on machine learning and perform repeated training and real-time recognition. For example, artificial intelligence can monitor whether a borrower's relevant device has been registered with multiple lending websites, whether the same device has downloaded multiple credit apps, whether it has initiated loan applications on multiple Internet credit platforms, etc., so as to reduce the likelihood of fraud. In addition, artificial intelligence is based on a huge knowledge graph and can also monitor information activities on the entire Internet. When it discovers that personal identifiable information has been leaked, it will call the police for you, which could be another great help in anti-fraud.

5.2.3 *Robo-advisory*

Robo-advisor, i.e., intelligent investment advisor, is another major application of artificial intelligence in the field of Internet finance. Artificial intelligence uses its deep learning system and network graph to deeply mine the captured user information and user behavior, and then accurately depicts user portraits from different dimensions such as occupation, income, risk preference, social class, and personality characteristics, thereby providing customers with personalized investment plans and suggestions. Robo-advisory is of great significance to financial institutions, such as commercial banks, third-party financial companies, asset management companies, and trusts. Taking insurance as an example, an intelligent investment consultant can calculate the insurance policy suitable for the customer through model analysis based on the customer's insurance experience, age, gender, income level, asset allocation, investment status, interests and hobbies, social relations, other detailed characteristic values, and product portfolio and provide reasonable insurance recommendations accordingly. Additionally, it gives customers who are frustrated with insurance products promoted by marketers a new channel to choose insurance plans. On the other hand, intelligent recommendations not only improve the user experience but also increase the policy closing ratio. Furthermore, to a certain extent, the labor costs are also reduced. For another example, the Turing product developed by InnoTREE is a typical case of robo-advisory. Through artificial intelligence technology, it can answer any investment and financing questions of users within three seconds and can automatically generate information including competing products and companies, industry, and other comprehensive data analysis reports.

5.2.4 *Intelligent customer service*

Intelligent customer service is another widespread application of artificial intelligence. Intelligent customer service uses machine learning and deep learning, text recognition, and voice data mining technology to automatically guess various questions that users may ask and uses the question recognition model to extract appropriate answers based on the user's usage scenarios and user behavior and collect the results, repeatedly

checking for accuracy. Take Ant Financial as an example. As early as the Double 11 Shopping Festival in 2015, 95% of Ant Financial's remote customer services were completed by big data intelligent robots, and 100% automatic voice recognition was achieved. Intelligent customer service can greatly reduce labor costs. In the Internet finance industry, and even in various other industries, customer service is an indispensable and huge labor investment, especially for those companies involved in marketing and consulting services. The application of intelligent customer service not only allows these enterprises and institutions to save labor costs but also provides customers with accurate, high-quality, and personalized services, which can improve customer satisfaction and provide customers with a better user experience. The application of artificial intelligence in the field of Internet finance, in addition to those described above, also includes automated branch services, credit financing, insurance pricing, core area security, and behavioral monitoring, etc. In the future, artificial intelligence will have wider ranges of applications and promote innovation and development of the Internet financial industry.

5.3 Internet Financial Risk Control and Deep Learning

As showed in the previous analysis, artificial intelligence technology is widely used in the field of Internet finance. The specific risk control issues of Internet finance include two key themes: risk control and anti-fraud. The use of deep learning technology can make up for the shortcomings of risk control methods in the traditional financial industry.

Taking consumer credit risk assessment in the field of risk control as an example, this problem can essentially be regarded as a classification problem, i.e., customers are divided into two categories: default and on-time repayment. The development of consumer credit risk prediction models has gone through three stages: qualitative analysis, statistical methods, and artificial intelligence methods. Qualitative analysis was the earliest method used for credit evaluation, and then statistical methods were gradually introduced into credit evaluation. In recent years, with the development of machine learning, some intelligent methods have

been gradually applied to credit assessment research. For example, Malekipirbazari *et al*., introduced in Chapter 4, used the random forest algorithm to predict the risk of borrowers on the online lending platform, Lending Club. However, the predictive effect of these traditional machine learning methods is very dependent on artificially designed features, and artificially designed feature methods often cannot take all features into account. At the same time, artificially designed features require a lot of time and labor costs. In recent years, deep learning has attracted more and more attention from scholars. Convolutional Neural Network (CNN) is one of the most classic and widely used network structures.

5.4 Application of Deep Learning in Credit Risk Control

In risk control scenarios, there is a large amount of time series data waiting to be mined in depth. For this type of data, the traditional method is to first construct the characteristics of the time dimension and then inputs them into the model together with other structured data for learning. Since the model fusion process should avoid being too complicated, this type of model is usually used for pre-fusion of generalized linear models. The most commonly used logistic regression algorithm in the field of risk control is essentially a generalized linear model, so there is the following widely used modeling method: First, use a model specifically for modeling time series data and output the predicted score (or hidden layer parameters), and then use the scores (or hidden layer parameters) as part of the features of the logistic regression scorecard.

Regarding time series data modeling, this section introduces a method of automatically learning features through the recurrent neural network (RNN). What performs better in practice is a special RNN model called long short-term memory (LSTM). This section first introduces the RNN model and then gradually transitions to the LSTM model.

5.4.1 *Relevant theories on credit risk control*

The LSTM neural network model is a type of RNN. RNN is a neural network specially designed to process time series data. It can infer current information based on historical information that appeared before. However, this model can only process information within a certain temporal range interval. If the time interval is too wide, it cannot accurately infer the information. On the other hand, LSTM can learn both long-term and short-term dependencies in time-series data, making them well suited for predicting time intervals and delayed events in time series.

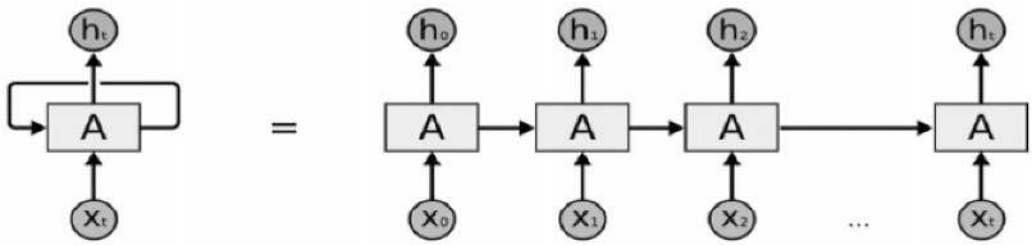


Figure 5.1. Recurrent neural network workflow diagram.

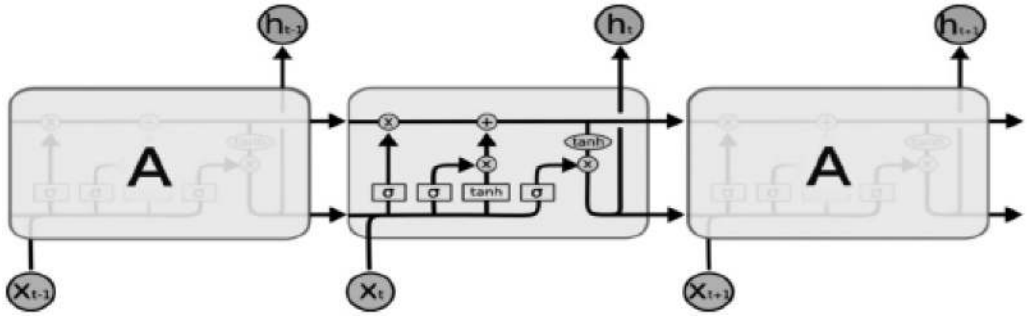


Figure 5.2. LSTM neural network workflow diagram.

An RNN can be seen as multiple copies of the same neural network. Each neural network module passes information to the next, as shown in Figure 5.1.

The biggest difference between the LSTM model and the RNN is the introduction of a dedicated cell state at the top of the LSTM, where the information is memorized. There are three control gates that control

whether the information is memorized: the input gate, the output gate, and the forget gate. Each neuron contains four interactive layers, as shown in Figure 5.2: the forget gate layer, the input gate layer, the input modulation gate layer, and the output gate layer.

The forget gate will read the output h_{t-1} from the prior time step and the current input x_t according to the formula

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + b_f\right) \tag{5.1}$$

Calculate the numerical value f_t and assign its value to the current cell state C_{t-1}.

The input gate layer consists of two parts: The first part determines the input value through the sigmoid function, and the second part creates a new candidate vector through the tan$h$ function ${C_t}^1$ and adds it to the state C_t. The calculation formula for this process is as follows:

$$i_t = \sigma\left(W_i \times [h_{t-1}, x_t] + b_i\right) \tag{5.2}$$

$${C_t}^1 = \tanh\left(W_c \times [h_{t-1}, x_t] + b_c\right) \tag{5.3}$$

The internal cell state C_{t-1} is then updated to C_t, and the calculation formula is

$$C_t = f_t \times C_{t-1} + i_t \times {C_t}^1 \tag{5.4}$$

The output gate layer determines which portion of the cell state that will be output by running a sigmoid function. It then multiplies the tan$h$-processed cell state by the output of the sigmoid gate to finally determine the output. This is calculated as follows:

$$o_t = \sigma_t\left(W_o \times [h_{t-1}, x_t] + b_o\right) \tag{5.5}$$

$$h_t = o_t \times \tanh(C_t) \tag{5.6}$$

5.4.2 *Research related to credit risk control*

In the traditional lending market, commercial banks occupy the most important market position. However, due to the cumbersome lending procedures of traditional commercial banks, strict requirements for borrowers, and financing costs, borrowers' financing needs have not been fully met or satisfied. The P2P lending industry developed earlier outside China than within China and the market is relatively mature. Research by Collier *et al*. (2010) shows that due to information asymmetry, banks are unable to accurately understand the borrower's creditworthiness and repayment ability and, as a result, will increase the overall borrowing interest rate, resulting in high financing costs for borrowers.[45]

In their research on the problem of financing difficulties, Agarwal and Hauswald (2010) found that the borrower's credit rating is the main factor limiting the success of their loans. In the online lending market, the credit rating problem can be solved, thus attracting the participation of a large number of borrowers.[46] Meyer *et al*.'s research (2007) shows that the default risk faced by P2P online lending platforms is due to (1) the information asymmetry between lenders and borrowers, which results in lenders not fully understanding the risks of the borrower, and (2) there is no collateral in place to guarantee the loan. Therefore, P2P platforms play an important role in the lending behavior.[47] Klafft (2008) found that when borrowing in an anonymous mode, the borrower's default risk will be greatly increased.[48] Research by Chaffee and Rapp (2012) shows that the emergence of a new lending model, such as P2P online lending platforms, caters to the financial needs of many individuals who have low credit qualifications and are unable to obtain sufficient financing from traditional credit institutions. In this environment, P2P platforms take advantage of investors' desire for high-profit investment and develop rapidly by taking advantage of legal loopholes when supervision and policies are not yet complete. Therefore, the platform faces extremely high risks of default. At the same time, there is also an urgent need for relevant departments to strengthen the regulation and to promote the healthy development of the industry.[49] Michels (2012) analyzed the role of online lending platforms in the entire lending process. Through a study of Prosper data, he found that lending platforms have a significant impact on the efficiency and risk of lending behavior.[50]

In the Internet era, the volume and growth rate of data are unprecedented. Due to the diversity of its data structure, big data includes massive structured data, semi-structured data, and unstructured data. All data must first be converted into a unified standard format before it can be utilized and its value mined. Big data has also been widely used as a risk control tool in the financial industry. Through the analysis of various transaction data accumulated on the Internet platform, cloud computing methods and models are used to help financial institutions prevent financial risks and strengthen risk control. In order to further understand the application of big data technology in the unique risks of Chinese P2P lending platforms, Xu *et al.* (2015) conducted further research on the risks, types, and risk identification of Chinese P2P lending behaviors.[51] In 2014, Kang and Cao studied the application of big data in the financial field and proposed the construction of a credit reporting system based on big data.[52] Other research on the application of big data in related industries is also gradually carried out.

The neural network simulates the transmission mode of signals between neurons in the human brain. When a neuron receives a signal, it activates the neurons in the hidden layer, and the signal is transmitted between the neurons, and then the corresponding response is made. Due to the differences in the basic structure of neural networks and the way of information transmission, there are also differences between neural networks.[53] West's research (2000) shows that traditional credit evaluation models rely on financial indicators and cannot reflect changes in evaluation objects in a timely manner. The logic and nonlinear analysis of neural network technology enable it to be flexibly applied to the credit evaluation system, and through simulation, the neural network in the human brain can find out the implicit connections between data so as to conduct credit evaluation more comprehensively and accurately.[54]

China's risk control research based on deep learning is still in its early stage. Dong Shen and Wang Qian (2016) proposed a group headquarters credit rating study based on the BP neural network artificial intelligence classification algorithm and learned the group's internal credit rating model through a supervised artificial intelligence algorithm.[55] Qiao Pengcheng and Zhu Junwei (2017) proposed the concept of artificial intelligence-based accounting credit system construction in view of the current situation of China's accounting credit system in the era of artificial

intelligence and the harm of the lack of a credit system.[56] Based on the hybrid fruit fly neural network, Wu Bin *et al*. (2017) proposed an improved BP neural network algorithm, which effectively improved the convergence speed and accuracy of the neural network and was used to build a P2P online loan personal credit risk assessment model, verifying that it has better learning ability and predictive ability.[57] Li Conggang and others (2015) conducted an in-depth analysis of the traditional financial and information technology risks faced by P2P online lending platforms, identifying current root causes of risks of online lending platforms, such as complex causes, serious uncertainty, and high externality, and proposed a BP neural network as a basic credit risk assessment model, and the validity and accuracy of the model were verified through MATLAB simulation.[58]

In credit risk modeling research, neural network algorithms have attracted the attention of researchers because they have certain self-learning capabilities and are good at handling complex nonlinear problems. There is also a wealth of research on its risk control applications for P2P network platforms. However, although there is a lot of research done on risk control based on neural networks, almost all of them is based on traditional shallow neural network models. Sirignano *et al*. (2016) have tried to use deep networks to evaluate mortgage risks with 294 explanatory variables.[59] However, they used multi-layer perceptrons instead of models such as CNNs. Most of the variables are loan-level characteristics and performance variables.

Although the concept of deep learning was proposed in the 1990s, it has not been taken seriously due to the technical conditions and limited use cases at that time. Deep learning has since developed rapidly with the rise of artificial intelligence and is widely used in all walks of life. Deep learning technology initially achieved good research results in fields such as computer vision and image recognition. In view of its powerful analytical capabilities in deep-level data correlation and considering the increasing complexity of Internet financial data in the current big data context, deep learning technology may also be applied to Internet financial risk control, credit assessment, and other fields.

The application of CNNs has gradually expanded to many fields including Internet financial risk control. Wang Chongren and Han Dongmei (2017) proposed a customer default risk prediction method based on CNN on the basis of dividing the input data into dynamic data and static data.

They compared this method with other machine learning algorithms and obtained better performance.[60] Wang Yang (2018) proposed a credit risk control analysis model based on CNNs to solve the problems of objectivity and hidden correlation discovery faced by traditional analysis models, organize data in time series, and build a set of deep learning credit assessment models. Empirical results show that deep learning models have higher accuracy than machine learning models, which also means that deep learning models have more advantages when analyzing complex nonlinear objects with multiple features.[61]

As a deep learning model, neural networks can acquire deeper and more generalizable features compared to artificially designed features. This is because neural networks have the ability to learn features hierarchically. Therefore, using CNNs to extract and analyze the deep connections between information can have a broad positive impact on the entire field of artificial intelligence. However, how to integrate CNNs into a new application model is a difficulty in current research, and there are few existing studies.

Therefore, if a risk identification model for borrowers on P2P online lending platforms can be established to address this difficulty, the P2P platform can measure borrowers' risks more accurately and comprehensively, and enhance the risk control capabilities of the P2P online lending platform.

5.5 Application of Deep Learning in Fraud Identification System

5.5.1 *Theories related to individual fraud*

Artificial Neural Network (ANN) itself has many shortcomings, such as too many training parameters and a complex network structure, which increases the difficulty of training. The existence of the above shortcomings became an important factor that hindered the application and development of ANN. It was not until neurology researchers made new discoveries in the cat brain in the 1960s that the neurons in the front of the visual system respond most strongly to specific light patterns. This discovery played a crucial role in the later creation of the concept of local receptive fields. On this basis, CNNs

began to appear and, to a large extent, made up for the shortcomings of traditional neural networks. Through an in-depth study of the working principle of the visual system of the animal brain, CNN simulates this, realizes feature extraction of original data, and improves the comprehensiveness and effectiveness of information extraction. As CNNs have the characteristics of sparse connections and weight sharing, the convolutional layer is responsible for feature extraction in the CNN structure, while the pooling layer implements dimensionality reduction of data. The two alternately appear, which not only effectively reduces the size of parameters but also completes the task of learning the characteristics of original data. As a typical deep learning model, CNN has the ability to learn features hierarchically. The essence of CNN is a deep neural network containing multiple hidden layers. Its algorithm ideas and training procedures are the same as those of traditional multi-layer perception machines. CNNs are different from deep neural networks in terms of the following:

(1) The neurons in each layer are not fully connected.
(2) The values of the local receptive field and the weight matrix during different convolution kernel operations are the same.

Due to the above properties of the CNN, compared with other deep learning algorithms, it reduces the number of training parameters and the complexity of the network structure to a certain extent, achieving accurate and efficient identification of abnormal trading behavior.

5.5.2 *Fraud related research*

Artificial intelligence technology represented by deep learning, combined with massive transaction data in a big data environment, can solve the difficult problem of high-dimensional complex data analysis, significantly reduce related labor costs, and strengthen its risk control and business processing capabilities. Effective applications have been achieved in the fields of intelligent finance and big data risk control. Compared with the technology of manual description and construction of features, deep learning can make full use of massive data resources in the context of big data to train models, describe the rich intrinsic information hidden in the

original data, obtain the most effective, and construct a relatively complex network structure to fully explore the correlation between data.

How to quickly and accurately identify fraudulent transactions is a huge challenge faced by the financial industry. Properly controlling fraud risks in finance, preventing illegal financial behaviors, avoiding economic losses, and preventing and controlling users' overconsumption and overindebtedness through technical capabilities are of great theoretical significance to the competitiveness and risk management capabilities of financial institutions and have significant practical value.

Different from the credit risk system, the fraud identification system solves a binary-classification problem, which is used to determine whether a user has fraudulent transaction behavior. The determination is mainly based on the user's basic information, transaction information, social information, etc. Labeling can reduce the huge losses to financial institutions caused by fraudulent transactions. In addition to commonly used classic machine learning algorithms, fraud identification systems will also involve knowledge in the field of deep learning. Based on the research on Bagging and Boosting, West *et al*. (2005) constructed an artificial neural network as a meta-learner, and then applied ensemble learning to it. Empirical research proved that the classification performance of the ensemble model was superior to that of a single model.[62]

In 2007, Kirkos *et al*. used neural network algorithms to identify fraud in Greek manufacturing companies and compared them with statistical methods; in 2011, Bhattacharyya *et al*. comparatively analyzed the logistic regression algorithm, support vector machine algorithm, and random forest algorithm in identifying credit card fraudulent transactions.[63] The results show that the logistic regression algorithm is inferior to the other two algorithms in terms of specificity and accuracy.[35] There are still many studies in the financial field based on ensemble learning. In 2015, Ala'raj and Abbod built a credit-scoring model based on homogeneous and heterogeneous classifier ensembles based on three base classifiers: LR, ANN, and SVM. Empirical research showed that integrating different base classifiers outperformed single classifier performance in terms of classification accuracy.[64]

Research on credit default identification in China has largely focused on the construction of credit risk scoring models, and there is a lack of research on the prevention and control of fraudulent transactions. In 2010, Wu

Baohua conducted in-depth research on the credit card anti-fraud system[65] and proposed a model based on an improved BP neural network with weighted decay. Given the extreme imbalance of credit card data, Liu Yanhong proposed to combine *K*-means clustering with the existing improved support vector data description (SVDD) algorithm to build a new single value classification model KmD-SVDD, and the ant colony algorithm is used to optimize the parameters of the model to improve the classification accuracy of the model. This fraud identification model divides users with similar transaction behaviors into the same category and then trains this type of transaction data. The final experimental results confirm that this model performs better in terms of its generalization ability.[66] In 2015, Ding Weixing[67] applied deep belief network (DBN) to the detection of fraudulent credit card transactions for the first time, providing reference value for the application of deep learning in this field in China. Based on the current difficulties in credit card overdue risk prediction and the excellent feature learning capabilities of deep learning, Xu Chao (2016) proposed a deep learning-based method for credit card overdue risk prediction and used multiple solutions to deal with the overfitting problem.[68] Fu Kang (2017) characterized a large amount of transaction data into the form of feature matrices and then applied CNNs to extract high-order features. At the same time, in order to make better use of a large amount of normal transaction data, the bagging integration method is used to train multiple CNN models and the predictive results of these models are averaged to obtain the final fraud score. Empirical research results show that the method of fusing multiple classifiers can significantly improve the robustness of the model.[69]

In fraud identification, the sample imbalance problem is a key factor affecting the performance of the model. This is due to the characteristics of user transaction behavior. A large amount of transaction data contains very small amounts of fraudulent transaction information, and labeling the data will consume large amounts of time and labor costs, so the small number of available labeled samples makes it difficult to meet the data scale requirements of deep learning. Wen Lina (2019) proposed a model based on semi-supervised deep learning[70] to address the problem of a lack of annotated data for real-time massive transaction data. Guanyu Chen (2018) integrated oversampling technology into the fraud detection model and

constructed a credit card fraud risk detection network model based on kNN-SMOTE-LSTM, which overcame the blindness and limitations of the SMOTE algorithm when generating new samples. Empirical research based on real credit card historical transaction data shows that this model exhibits superior Internet financial fraud detection performance and provides a certain theoretical basis and practical reference for financial institutions to apply deep learning technology for Internet financial risk detection.[71] Chen Rongrong (2019) proposed a fraud detection fusion model of generative adversarial network and extreme gradient boosting machine and used real historical transaction datasets to conduct comparative experiments, thus confirming the feasibility and effectiveness of the model.[72]

With the development of Internet finance and the advent of the big data era, the number and amount of loan transactions have increased significantly, and risk management is of great significance. Traditional machine learning models encounter bottlenecks in training and prediction under massive, high-dimensional, and complex data. In recent years, the emergence of deep learning technology has solved the defects of shallow neural networks and can construct multi-layer nonlinear relationships, which has been widely used in many fields. It is of practical significance to reduce credit default risks, strengthen financial supervision, and reduce losses of financial institutions through methods based on deep learning.

Chapter 6

Application of Graph Model in Research on Internet Financial Risks

6.1 Introduction

In recent years, the widespread adoption of Internet technology has given rise to a novel financial business model known as Internet finance. Its rapid development is attributed to its efficiency, cost-effectiveness, diverse service offerings, and inclusive nature. However, as Internet finance expands, associated risks, particularly those related to fraud, have become increasingly apparent. Consequently, there is an urgent need to establish an Internet financial fraud identification system.

Individuals' behaviors in society are interconnected, and a comprehensive understanding of their true situation can be understood through their social networks. The same principle applies to fraud detection. When determining whether an individual has engaged in fraudulent activities, analyzing their social relationships and environment, in addition to their personal information, enhances the effectiveness of fraud detection. The examination of a person's social network requires the use of a graphical model.

A graphical model is a representation of a system composed of points and lines, forming a graph that describes relationships between numerous entities in nature and human society. This chapter primarily explains the application of graphical models in Internet financial fraud detection.

6.2 Graph Model Theory

A graph is a format of linked data where graphical networks can represent a variety of complex systems. A graph is defined by nodes and edges. The mathematical expression of nodes is denoted as $V = \{v_1, v_1, \ldots, v_1\}$, and the mathematical expression of edges is denoted as: $E = \{e_1, e_2, \ldots, e_m\}$. The composition of a graph can be represented by this tuple $G(V, E)$.

In a graph, nodes represent entities, and edges denote relationships between these entities. For instance, in social contexts, nodes may represent social entities such as individuals, while edges illustrate social relationships between people.

Graphs can be categorized based on the properties of edges into undirected graphs and directed graphs. That is, if the edges $e(v_1, v_2)$ and $e(v_2, v_1)$ represent the same relationship, it is called an undirected graph; if the edges $e(v_1, v_2)$ and $e(v_2, v_1)$ represent different relationships, it is called a directed graph, as shown in Figure 6.1.[73]

Network structures associated with fraud can be classified into two categories: node classification problems and community mining problems.

Node Classification Problem: In the network, every entity (borrower) within the network is the object of analysis. The objective of fraud detection in this context is to assess the possibility of fraud for each individual node.

Community Mining Problem[74]: On the Internet, certain fraud groups exhibit well-defined divisions of labor and employ interlocking fraudulent tactics. To address such fraud identification challenges, the community mining method of graph networks can be used to estimate the collective fraud risk.

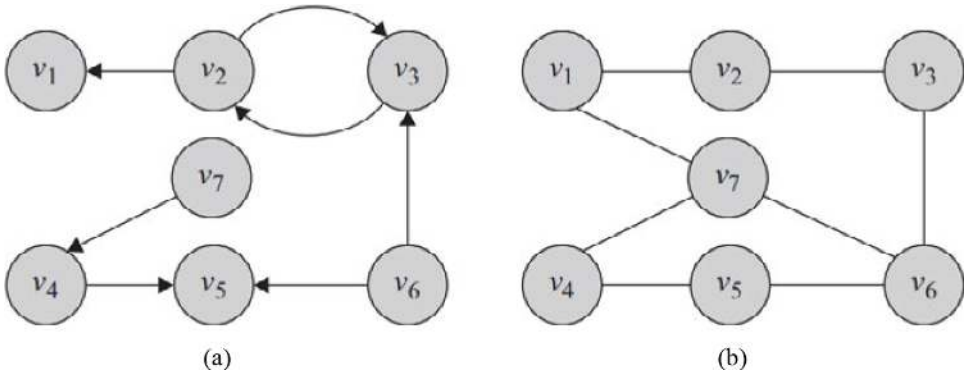


Figure 6.1. Directed and undirected graphs. (a) Directed graph and (b) undirected graph.

6.2.1 *Node classification*

As showed in Figure 6.2, some nodes have observed labels, while the remaining nodes have established connections with these labeled nodes, signifying that they are linked to the labeled nodes through edges, yet their labels remain unobserved. The node classification problem is to determine the appropriate labels for a node based on its neighbors' labels and the relationships between nodes. From the risk control perspective, analyzing this network allows for the calculation of the probability distribution of related nodes engaging in fraudulent behavior after specific nodes in the network do so.

From a purely graphical point of view, the node classification problem can be categorized into two methods: naive classification and collective classification.

6.2.1.1 *Naive classification*

The fundamental steps of naive classification method are outlined below:

(1) Derive the adjacency matrix from graphs;
(2) Model the adjacency matrix and node attributes using machine learning algorithms; and
(3) Assign labels to the unlabeled node.

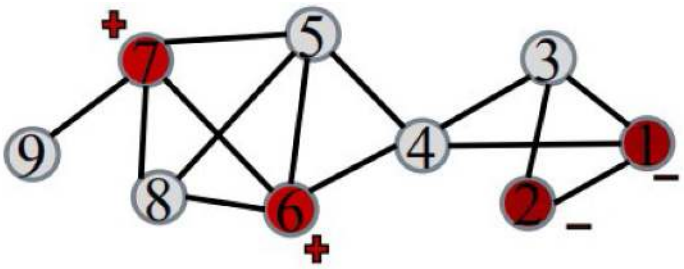


Figure 6.2. An example diagram of node classification.

Taking Figure 6.2 as an example, the information matrix of every node (adjacency matrix + node attributes) is first constructed to characterize such node.

$$u_1 = [0,1,1,1,0,0,0,0,0] \qquad y_1 = 0$$

$$u_2 = [1,0,1,0,0,0,0,0,0] \qquad y_2 = 0$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$u_7 = [0,0,0,0,1,1,0,1,1] \qquad y_7 = 0$$

The above information is then classified using a classifier within a machine learning framework, and the outputs are subsequently used to compute the remaining nodes.

6.2.1.2 *Collective classification*

The naïve classification method considers the static influence of adjacency information, while the collective classification method considers a dynamic influence process.[77] This dynamism is evident in the process that each node disperses its own information. Specifically, each node first updates its information upon receiving new data and subsequently transmits the updated information to its neighbors, constituting an iterative message-passing algorithm. The primary steps of this method are outlined below:

(1) Simulate the graph propagation process and compute the graph;
(2) Evaluate the degree of propagation impact on unlabeled nodes; and

(3) Assign probability to unlabeled nodes.

6.2.1.3 *Commonly used algorithms*

(1) Weighted Average

The weighted average method considers the class probability of Y_i to be the weighted average of the class probabilities of its neighbors.[75]

$$P(Y_i = c) = \frac{1}{Z} \sum_{(Y_i, Y_j) \in E} w(Y_i, Y_j) P(Y_j = c) \tag{6.1}$$

The algorithm simply requires iterating the above formula until convergence. To some extent, the weighted average method is also a description of the information dissemination process, in which the weight represents the degree of acceptance in the information dissemination process.

The disadvantage of this method is that the formula does not consider the properties of the node itself but only calculates a weighted average value of the labels of the neighbor nodes, which is not universal.

(2) Iterative Classifier

The dynamics of the iterative classifier are reflected in[76]:

(i) First train a classifier based on observed labels $f(\bullet)$.
(ii) According to the trained classifier, update and iterate the labels of unlabeled nodes and dynamically update the attributes of each node.
(iii) The change of the unlabeled node dynamically updates the attributes of other adjacent nodes and a new classification is obtained.

Through this dynamic iteration, the adjacency information is continuously updated, making this iterative classification method more effective than the simple static classification method.

(3) Loopy Belief Propagation

The impact of a labeled y on all unlabeled nodes Y can be expressed as:

$$P(y|x) = \frac{1}{Z(x)} \prod_{Y_i \in y} \Phi_i(y_i) \prod_{(Y_i, Y_j) \in E} \Psi_{ij}(y_i, y_j) \tag{6.2}$$

$$\Phi_i(y_i) = \Psi_i(y_i) \prod_{(Y_i, X_j) \in E} \Psi_{ij}(y_i) \tag{6.3}$$

The intuitive explanation of the above formula is as follows:

First, $P(y|x)$ represents the joint distribution of the graph potential, treating each node label Y_i as a random variable.

Second, $\Phi_i(y_i)$ represents a "known" potential, indicating the ability to disseminate information.

Third, $\Psi_{ij}(\bullet)\Psi_i(\bullet)$ is a potential function.

$\Psi_i(y_i)$ represents a priori potential, $\Psi_{ij}(y_i)$ represents the potential of observed variables on nodes, and $\Psi_{ij}(y_i, y_j)$ represents the potential between nodes. This allows us to define a pairwise Markov random field as a joint distribution of unobserved labels. $\Psi(\bullet)$ is often defined as an exponential function.

$$\Psi(x) = e^{-H(x)} \tag{6.4}$$

$$H(x) = \sum_{u,v \in x, i=v} \alpha_{uv} x_u x_v + \sum_{v \in x} \beta_v x_v \tag{6.5}$$

Here, $\alpha_{u,v}$ and β_v are the variables that need to be learned.

Define the message sent from node i to node j as $m_{i \to j}(y_j)$, and then:

$$\Phi_i(y_i) = \Psi_i(y_i) \prod_{(Y_i, X_j) \in E} \Psi_{ij}(y_i) \tag{6.6}$$

$$m_{i \to j}(y_j) = \alpha \sum_{y_i \in \xi} \Psi_{ij}(y_i, y_j)\, \Phi_i(y_i) \prod_{Y_k \in N_i \cap y \setminus Y_j} m_{k \to i}(y_i) \tag{6.7}$$

The iterative classification and loopy belief propagation algorithms can be written in the form of Figure 6.3. For further details, refer to the provided reference.[77]

6.2.2 *Community mining*

Many complex systems in the natural and social domains can be formulated as networks consisting of sets of nodes connected by edges, such as the Internet, communication and distributed networks, logistics and supply chain networks, industrial cluster networks, and social organization networks in the real world. And it has been found that an important feature exists in most of the real networks — community structure, i.e., the whole network is composed of several communities. These communities have the structural characteristics of tightness inside and looseness outside, that is to say, the connection of nodes inside the community is very tight, while the connection of nodes between the communities is relatively loose. Communities are also called "clusters," and the nodes in a community usually share similarities.

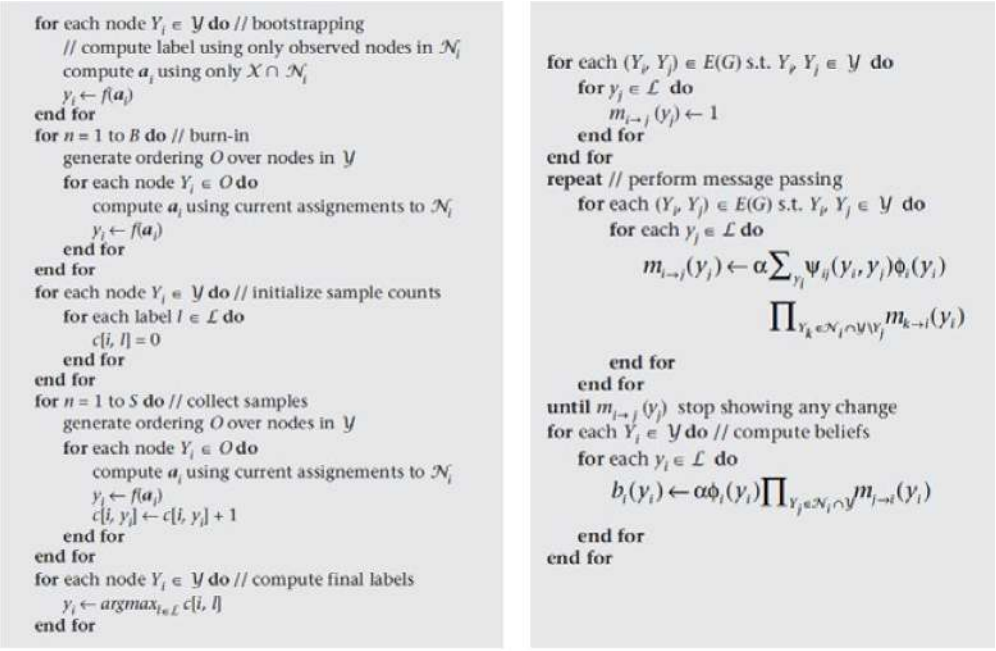


Figure 6.3. (a) Iterative classification algorithm; (b) Loopy belief propagation algorithm.

6.2.2.1 *Research on community mining in static networks*

Community mining is actually a network clustering method. Since there is no clear concept and definition of community, scholars have also proposed community detection algorithms based on objective understanding, and each algorithm has its own defined assumptions and interpretations of network community structure.

The community detection problem is closely related to graph clustering, and most of the traditional detection methods are accomplished with the help of graph clustering or graph partitioning. Girvan and Newman were the first to propose the concept of modularity, and transformed the algorithms for community discovery into the problem of solving the community partitioning and maximizing the network modularity.[78] Although this density-based community definition has practical significance in most networks, there are still some networks that cannot be divided into communities in this way. For example, in Twitter's user-following network, politicians, actors, singers, and these groups with a large number of followers can be regarded as different communities, while their respective fan groups, such as users who tend to follow politicians, because of the existence of common preferences, can also be considered as a community. The division of communities is not based on the principle of dense connections within communities and sparse connections between them but on the fact that the nodes within a community share a common "connection pattern." For example, in the example above, the fan base of politicians tends to follow politicians, and this common "connection pattern" reflects the similarity between the nodes within the community.

In addition, according to whether a node is allowed to belong to more than one community at the same time, communities can be categorized into overlapping and non-overlapping communities. The so-called "overlapping" refers to the fact that a node can belong to more than one community at the same time, which is very common in real life. For example, we play different roles in our family, work, and life, so we can be in multiple communities at the same time. Allowing a node to belong to more than one community will complicate the model and algorithm, but it is also more realistic. In addition, the communities in the network may also have a certain hierarchical structure, for example, for a student, their class can be regarded as a community, and their school can also be regarded as a community, and the class is a subset of the school. This suggests that there

may be a hierarchical relationship between communities. Community discovery is further complicated by the overlapping and hierarchical community structures that may exist in a network.

6.2.2.2 *Research on community detection in dynamic networks*

In recent years, community detection in dynamic networks has garnered significant attention, and various methods addressing this challenge have been proposed. However, progress in this area has been limited, primarily due to two main factors. First, there remains considerable debate regarding community detection methods in static networks. Second, the acquisition of real-time network data continues to pose a significant challenge in the research process. Recently, with the release of certain dynamic network datasets, researchers have gained the capability to monitor the dynamic evolution process of systems. This newfound ability opens up opportunities to study the formation, evolution, and disintegration of community structures within networks.

6.2.2.3 *Overview of community detection methods*

Community detection methods can be divided into two categories: partition-based methods and label propagation-based methods.

(1) Method Based on Partition

The basic idea: Initially, identify all links between communities, remove them, and subsequently, each connected branch corresponds to a community. The central aspect of this method lies in finding links between communities and generating appropriate community divisions. One of the most well-known methods in this category is proposed by Girvan and Newman.[78] This method characterizes network connectivity through betweenness and divides the network based on maximum betweenness.

The basic steps are as follows:

(i) Calculate the betweenness of each edge in the network;
(ii) Remove the edge with the largest betweenness and recalculate the betweenness of the remaining edges; and
(iii) Repeat the above steps until no edges can be removed.

(2) Method Based on Label Propagation

This approach is a semi-supervised learning method based on graphs. The fundamental concept involves leveraging the label information of labeled nodes to predict the labels of unlabeled nodes.

Initially, a complete graph model is established using the sample relationships. In a complete graph, nodes include both labeled and unlabeled nodes, and edges represent the similarity between nodes. Node labels can be passed on to other nodes based on their similarity, akin to labeled data acting as a source for unlabeled data. The higher the similarity between nodes, the more effectively labels spread.

The core idea is that in a network with a community structure, any node should belong to the same community as most of its neighbors. Based on this concept, the well-known Label Propagation Algorithm (LPA) was introduced.[79] The algorithm's execution involves initializing each node with a unique label. In each iteration, each node uses the labels of its majority neighbors to update its own label. The algorithm concludes when the labels of all nodes align with the majority of their neighbors' labels.

While the algorithm is simple, easy to implement, exhibits a short execution time, and has good interpretability, its instability led to improvements in literature to ensure the convergence of the label propagation-based method.[80]

6.2.3 *Graph-based learning methods*

Graph-based learning methods mainly include graph representation models and graph neural network models.

Graph Representation Model: Studies how to jointly model a node's intrinsic features and graph features. Among them, the graph features are the description and characterization of the given graph. It mainly includes using the representation learning process in deep learning to characterize the nodes of the graph, to characterize the graph, and so on. Typical algorithms include DeepWalk, Node2Vec, and Struc2Vec.

Graph Neural Network Model: Studies how to design iterative algorithms for graph networks so that they can both incorporate flexible neural network structures and introduce neural networks' methods of training and

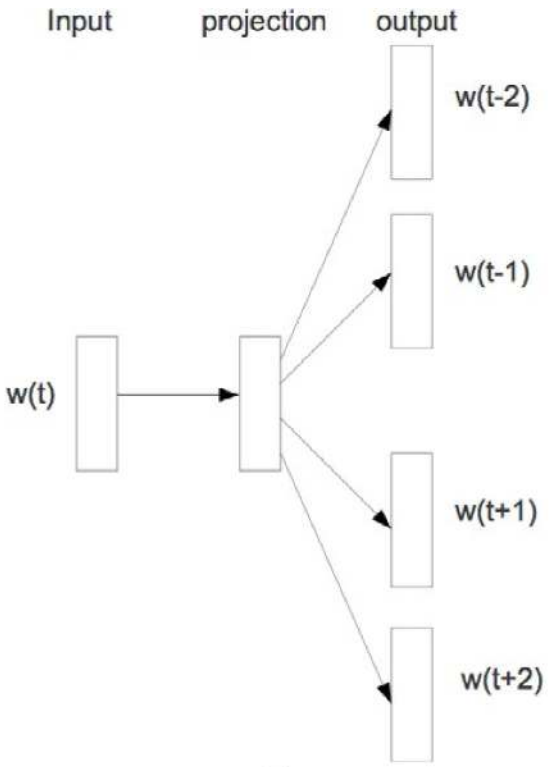
optimizing parameters. Typical algorithms include Semi-GCN, GraphSAGE, and GATs.

6.2.3.1 *Graph representation model and its typical algorithm*

The idea of graph representation modeling comes from the word embedding method in natural language processing. In natural language processing, Word2Vec is a method of converting words into vectors. The commonly used Word2Vec methods include CBOW and Skip-Gram algorithms. The following is an example of Skip-Gram to illustrate how to realize the vectorization of words.

As shown in Figure 6.4, the idea of the Skip-Gram algorithm is to predict the context words $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$ based on the given center word $w(t)$.

For specific details and algorithms of Skip-Gram, please refer to Word2vec Mathematical Principles or the original text.



(a)

$$\begin{aligned}\text{minimize } J &= -\log P(w_{c-m},\ldots,w_{c-1},w_{c+1},\ldots,w_{c+m}|w_c)\\ &= -\log \prod_{j=0,j\neq m}^{2m} P(w_{c-m+j}|w_c)\\ &= -\log \prod_{j=0,j\neq m}^{2m} P(u_{c-m+j}|v_c)\\ &= -\log \prod_{j=0,j\neq m}^{2m} \frac{\exp(u_{c-m+j}^T v_c)}{\sum_{k=1}^{|V|}\exp(u_k^T v_c)}\end{aligned}$$

(b)

Figure 6.4. Skip-Gram algorithm.

(1) DeepWalk

DeepWalk is essentially an application of the Skip-Gram algorithm to graph.[81] In the field of natural language, each sentence in the corpus can describe the co-occurrence relationship between words, and through this co-occurrence relationship, word embedding can be realized, thus enabling the word vectorization. In the field of graphs, a graph is analogous to the entire

corpus, and the nodes in the graph are analogous to words. The node embeddings can be realized through the co-occurrence relationship between nodes.

Description of the co-occurrence relationship between nodes: Given a node u_i, let u_i randomly walk on the edges of the graph to its neighbors and eventually get a random walk sequence of the node u_i. We can use the Skip-Gram algorithm to model this sequence. DeepWalk's method is to construct a sequence of "words" on the graph, and then train it directly with the Word2Vec model to obtain a representation vector for each node in the graph.

(2) Node2Vec Algorithm

The DeepWalk algorithm provides a sequence generation method based on depth-first search. In general, models based on breadth-first search tend to obtain content similarity (i.e., homophily) in graphs, while models based on depth-first search tend to obtain structural similarity. The Node2Vec algorithm attempts to provide a sequence generation method that can take into account both depth and breadth search, as showed in Figure 6.5.[82]

As showed in Figure 6.6, suppose we start a random walk from the node t and now reach the node v. Under a breadth-first-style walk, the next position should be the node x_1; under a depth-first-style walk, we would reach either the node x_2 or node x_3, with the possibility of returning to the node t.

In order to describe this randomness, two parameters *p and q* are added to control the sequence generation process.

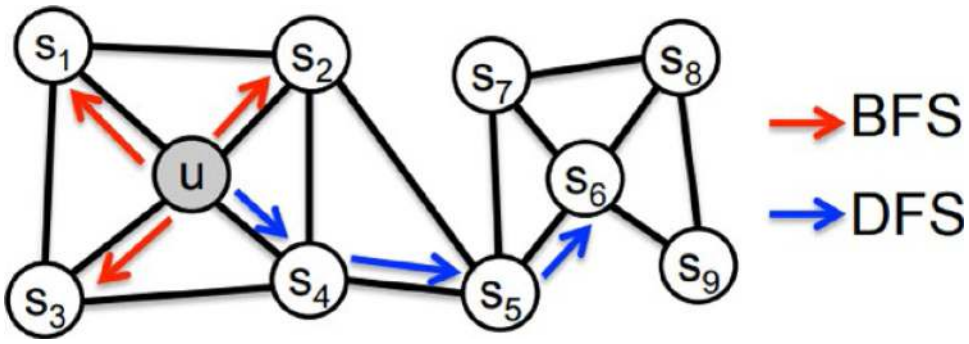


Figure 6.5. Node2Vec search.

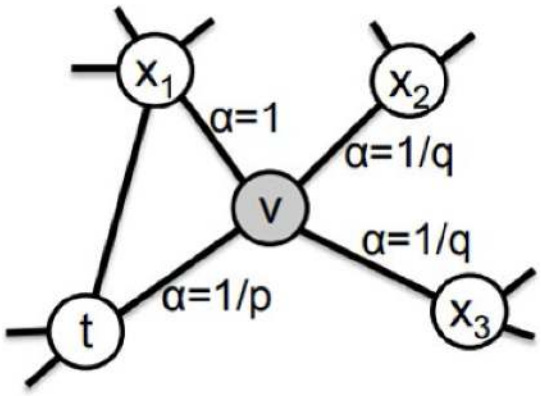


Figure 6.6. Node2Vec controlled searching.

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} \ if \ d_{tx} = 0 \\ 1 \ if \ d_{tx} = 1 \\ \frac{1}{q} \ if \ d_{tx} = 2 \end{cases} \tag{6.8}$$

where v is the current node, t is the previous step of v, x represents the position of the next step, and d_{tx} represents the shortest distance between t and x:

(i) When $d_{tx} = 0$, that is, from the node v back to the node t, the search parameter is $\frac{1}{p}$.
(ii) When $d_{tx} = 1$, that is, the node x is a neighbor of the node t, which is equivalent to breadth-first search, the search parameter is 1.
(iii) When $d_{tx} = 2$, that is, the node x is the neighbor of the node t, which is equivalent to depth-first search, the search parameter is $\frac{1}{q}$.

(3) Struc2Vec Algorithm

The Struc2Vec algorithm argues that node embeddings should not only consider the adjacency of nodes but also need to consider the structural similarity of nodes.[83]

As shown in Figure 6.7, the nodes u and v are very similar in structure to each other, but if there are numerous intermediate nodes and connections between them, it is difficult to characterize the nodes u and v as similar vectors no matter what path the random walk takes. Struc2Vec argues that

models such as DeepWalk and Node2Vec are effective because adjacent nodes in the real world often have relatively similar local neighborhoods.

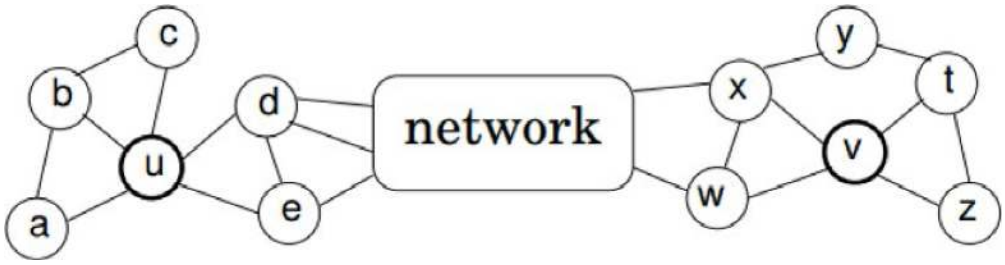


Figure 6.7. Similar structures in Struc2Vec.

One of the main innovations of the Struc2Vec algorithm is that when randomly walking through the graph, one does not only start walking from the neighbors of a node but also starts from nodes with similar structures. The node sequence generated in this way is a sequence that can well represent the structure, and then the representation vector of each node can be obtained by modeling with the Skip-Gram algorithm.

Ribeiro, Saverese, and Figueiredo defined an iterative approach to calculate the similarity between nodes[83]:

$$f_k(u, v) = f_{k-1}(u, v) + g\left(s(R_k(u)), s(R_k(v))\right) \tag{6.9}$$

where $R_k(u)$ denotes the node set where the shortest distance from the node u is k. $s(S)$ represents the orderedness of the set S, and $g(\bullet)$ is a dynamic time warping function used to calculate the distance between time series with mismatched lengths.

When $k = 1$, $f_1(u, v)$ describes the similarity between nodes u and v using neighboring nodes with a distance of 1. Similarly, when $k = 2$, $f_2(u, v)$ describes the similarity between nodes u and v using neighboring nodes with a distance of 2.

If a matrix consists of the values obtained by $f_k(u, v)$, the matrix is considered as an adjacency matrix, then, by means of construction of $f_k(\bullet)$, a hierarchical graph can be obtained.

Next, the series can be constructed through random walk. First, let each step have p probability of being in this layer, and $1 - p$ the probability of moving to a neighboring layer, and then define a search method similar to Node2Vec in each layer to construct the series.

Once the series that can represent the structure is obtained, representation learning can be performed on it to obtain the representation vector of each node. There are some more details of the strategy and optimization algorithm. Please refer to the literature,which will not be described in detail here.[83]

6.2.3.2 *Graph neural network model and its typical algorithm*

Graph representation is an approach where features are extracted and then learned using machine learning algorithms. We would like to build a model on the graph that can directly accomplish a given task, such as the fraudulent node identification problem.

The graph neural network model was first proposed in 2009, and its core idea is to define both node classification and message passing processes within an extensible neural network model, thus combining neural networks with graphs. However, the disadvantage of this method is that the computational complexity is too high, and there are a large number of iterative computations in the forward propagation process, which makes it difficult to compute the gradient in the backward propagation process. However, scholars have begun to improve this framework by incorporating computational modules that are commonly used in deep learning, such as convolutional neural network (CNN), recurrent neural network (RNN), and attention mechanisms (Attention), in order to improve the capabilities of graph neural network models.

In recent years, graph neural networks have become a powerful and practical tool for machine learning tasks in the graph domain. Literature categorizes existing graph models integrated with deep learning into three groups: semi-supervised models, unsupervised models, and some of the latest models.[84] Semi-supervised methods can be further categorized into graph neural networks and graph convolutional networks. Recent advancements include graph recurrent neural networks and graph reinforcement learning methods.

(1) Semi-GCN Algorithm

In real fraud problems, there are often a large number of unlabeled nodes, but since the graph structure can reflect the similarity between nodes, adding a large number of unlabeled samples (nodes) to the training model can improve the classification effect. Therefore, the proposal of graph-based semi-supervised learning method is very meaningful.

The semi-GCN method proposed in the literature is a third-generation graph convolution network method that uses the results of the formula of Chebyshev approximation of spectral convolution when $k = 1$ to improve the previous convolution.[84] Its convolution formula is given by:

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{6.10}$$

where $H^{(l)}$ is the convolution output of the previous layer, $\tilde{A} = A + I$, A is the adjacency matrix, and I is the identity matrix. $\widetilde{D}_{ii} = \Sigma_j \widetilde{A}_{ij}$ is an approximate matrix derived. $\sigma(\bullet)$ is the activation function.

Figure 6.8 is used as an example to explain the training process of Semi-GCN: Take two layers as an example:

$$Z = f(X, A) = softmax(\hat{A}ReLu(\hat{A}XW^{(0)})W^{(1)})$$

Calculate the node feature matrix $X \in R^{N\times C}$and adjacency matrix $A \in R^{N\times N}$, and preprocess the adjacency matrix $\hat{A} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$.

First layer of convolution + *ReLU* non-linear activation function: $H^{(0)} = ReLu(\hat{A}XW^{(0)})$

Second layer of convolution + *Softmax* activation function leads to the following output:

$$H^{(1)} = softmax(\hat{A}H^{(0)}W^{(1)}) = softmax(\hat{A}ReLU(\hat{A}XW^{(0)})W^{(1)})$$

$$\Rightarrow Z = H^{(1)}$$

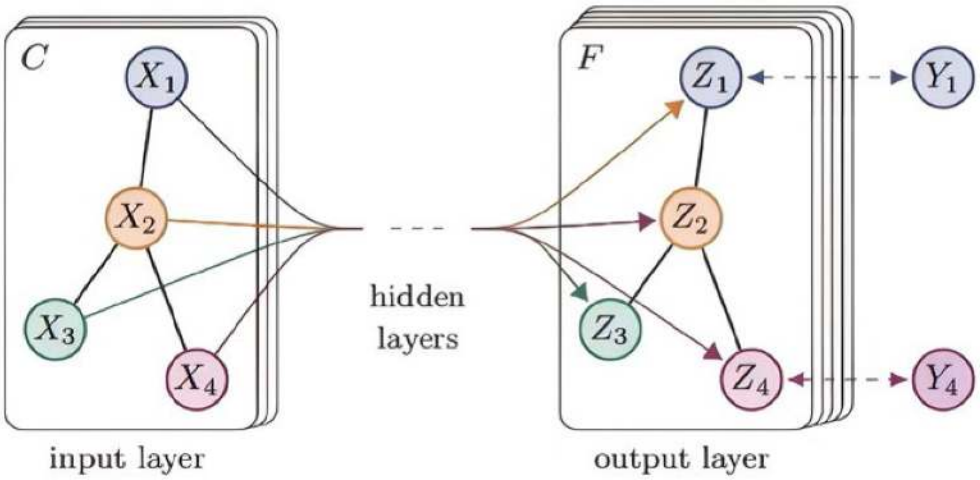


Figure 6.8. Semi-GCN example.

(2) GraphSAGE Algorithm

The method described above is a direct inference method, meaning that inference is performed on a fixed graph. If a new node appears in the graph, the entire graph will need to be retrained. Therefore, a more efficient training method needs to be explored.

The training process for the Semi-GCN algorithm involves calculating the cross-entropy loss of identified nodes. However, computing the information of these nodes as input to the neural network requires calculating the entire graph, which consumes significant computing power. To address this issue, the literature suggests using an inductive method.[85] This involves utilizing a function (convolutional layer) to train the aggregate feature information from neighboring nodes. This extends GCN into an inductive learning task, allowing it to generalize to unknown nodes.

The graphical process of the algorithm is shown in Figure 6.9:

(i) First, randomly sample the neighbors to reduce the complexity of the computation (the number of one-hop neighbor samples = 3 and the number of two-hop neighbor samples = 5).

(ii) Generate the embedding vector of the target node: first, aggregate the features of the 2-hop neighbors to generate the embedding vector of the one-hop neighbor and then aggregate the embedding vectors of the one-hop neighbor to generate the embedding vector of the target node.

The embedding vector is fed through a fully connected layer to predict (iii) the label of the target note.

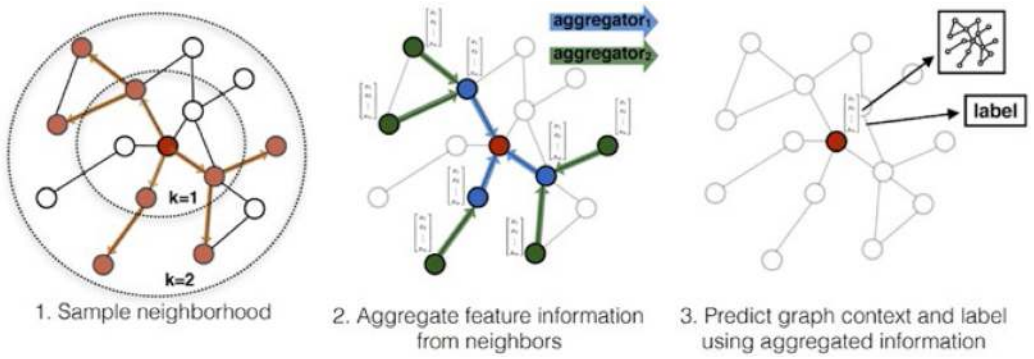


Figure 6.9. GraphSAGE example.

Literature proposed a number of aggregators, such as mean aggregator and pooling aggregator.[85] The basic idea is as follows:

First take the mean of each dimension in the neighbor embedding and then concatenate it with the embedding of the target note and perform non-linear transformation, or take the maximum value of the non-linear transformation in each dimension.

$$h_{N(v)}^{k} = mean\left(\left\{h_{u}^{k-1}, u \in N(v)\right\}\right) \tag{6.11}$$

$$h_{v}^{k} = \sigma\left(W^{k} \cdot CONCAT\left(h_{v}^{k-1}, h_{N(u)}^{k-1}\right)\right) \tag{6.12}$$

$$h_{N(v)}^{k} = \max\left(\left\{\sigma(W_{pool} h_{ui}^{k} + b)\right\}, \forall u_i \in N(v)\right) \tag{6.13}$$

$$h_{v}^{k} = \sigma\left(W^{k} \cdot CONCAT\left(h_{v}^{k-1}, h_{N(u)}^{k-1}\right)\right) \tag{6.14}$$

(3) GATs Algorithm

The GATs algorithm is actually an extension of the GraphSAGE algorithm, i.e., an aggregation function that uses the attention mechanism. In order to better express the innovation of the GATs algorithm, we first show the pseudocode of the GraphSAGE algorithm.

As can be seen from Figure 6.10, the GraphSAGE algorithm mainly integrates the aggregation process of the target point's neighbors and the estimation process of that point after aggregation.[85] The framework can be described as follows:

Algorithm 1: GraphSAGE embedding generation (i.e., forward propagation) algorithm

Input : Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$; input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$; depth K; weight matrices $\mathbf{W}^k, \forall k \in \{1, ..., K\}$; non-linearity σ; differentiable aggregator functions $\text{AGGREGATE}_k, \forall k \in \{1, ..., K\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

Output : Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

$\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$;
for $k = 1...K$ **do**
  for $v \in \mathcal{V}$ **do**
    $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
    $\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$
  end
  $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2, \forall v \in \mathcal{V}$
end
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$

Figure 6.10. GraphSAGE algorithm framework.

(i) Randomly sample the neighbors and retrogrades of the target node, and generate the neighboring features of the target node through an aggregation function.
(ii) The target node's feature vector is updated by combining its own features with those of its neighbors.
(iii) Estimation of the target nodes using neural networks.

GATs are improved by the use of the attention mechanism in step 4 in Figure 6.10 with an aggregation function of [86]:

$$h_{N(v)}^k = \sigma\left(\sum_{j \in N(v)} a_{ij} W h_j\right) \tag{6.15}$$

where a_{ij} is the attention factor, and W is a function of the transformation of the original features.

Due to the introduction of the attention mechanism, the algorithm process becomes complex.

Below is a high-level summary of the algorithm from the input to hidden layer to output.

Input Preprocessing: Perform feature transformation on feature vectors of nodes h_i, h_j and get Wh_i, Wh_j:

(i) W is a parameter matrix that maps features in the original dimensions to features in the fixed dimensions.
(ii) The purpose of linear transformation is to obtain vectors with stronger expressive power.

Input layer → Hidden layer: Calculate the importance between nodes:

$$e_{ij} = LeakyReLU(\alpha^T[Wh_i||Wh_j])$$

(i) $e_{i,j}$ denotes the importance of the features of the neighboring node j to the node i;
(ii) $[Wh_i||Wh_j]$ denotes the operation of concatenating vectors Wh_i with vectors Wh_j;
(iii) α is a shared parameter from the input layer to the hidden layer.

Hidden layer → Output layer: In order to make the contributions of different neighbors comparable, *softmax* normalization is used, and finally, the contribution of the neighboring node j to the generation of new features of the node i is obtained, which is denoted as

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{k\in N_i}\exp(e_{ik})} = \frac{\exp\left(LeakyReLU(\alpha^T[Wh_i||Wh_j])\right)}{\sum_{k=1}^{k\in N_i}\exp\left(LeakyReLU(\alpha^T[Wh_i||Wh_k])\right)}.$$

Output → Aggregation: The algorithm finally calculates a weighted average of the features of node i's neighbors by their contribution α_{ij}. It then applies a non-linear transformation to obtain the new features of node i:

$$h_{N(v)}^{k} = \sigma\left(\sum_{j \in N(v)} a_{ij} W h_j\right).$$

6.3 Application of Graph Model in the Field of Internet Financial Risk Assessment

Based on its specific fraud scenario application, we classify the literature into two categories. The main method of the first category is to extract features from graphs and use existing machine learning methods to train for fraud detection. We classify this method as a machine learning method. Another main method is to use graph models to identify fraud models. This type of method is different from feature extraction, and we categorize this approach as a deep learning method. There are also some other methods that use graph representation learning models to extract features and then train them in machine learning models for fraud identification. These methods still adopt the logic of machine learning models, but due to the innovative nature of their feature extraction, they are categorized as deep learning methods.

6.3.1 *Applications of graph models combined with machine learning*

Currently, most papers that use graph methods for fraud detection focus on node classification problems in networks. This method is relatively intuitive, and its basic idea is: first, some features about the graph are extracted with the node's own characteristics on the graph as the feature matrix of the node. Then, machine learning algorithms are used on this matrix for classification. The method based on graph features is currently the most commonly used method. The key to the effectiveness of this method lies in whether the features are effective or not. Therefore, for the problem of fraud detection, scholars have proposed various methods to find effective features.

6.3.1.1 *Feature extraction method based on graph statistics*

Complex networks themselves have many statistical properties, such as in-degree, out-degree, clustering coefficient, node betweenness, etc. Using these features, we can calculate a vector with graph statistical properties for each node of the network as a feature based on the network structure and merge it into the features of the node itself.

6.3.1.2 *Feature extraction method based on node influence*

Statistics-based methods are relatively simple and adaptive. But in order to find more useful features, a more common approach is to use the propagation of known labels on the graph to obtain some features that describe the importance of nodes. These features are combined with the original node features as feature vectors extracted from the graph to form a feature matrix.

There are two main algorithms: PageRank algorithm and Personalized PageRank algorithm:

Van Vlasselaer *et al*. used the Personalized PageRank algorithm for feature extraction.[91] There are many similar articles that use node influence algorithms to extract features. For example, the literature[88] uses the PageRank algorithm to score and identify fraudulent calls, and literature[89] uses the PageRank algorithm to greatly improve the accuracy of the model.

The main idea of the PageRank algorithm is to treat the graph of nodes as a link structure and propagate fraud information from labeled fradulent nodes to informational nodes and then to unlabeled nodes in proportion to the strength of the link connections.

The classic PageRank algorithm only considers the topology of the graph, and when it is placed in an anti-fraud scenario, it cannot effectively utilize the labeled data. And the extension of the PageRank algorithm — the Personalized PageRank algorithm, can effectively solve this problem. Taking web pages as an example, users randomly wander (switch) between web pages. In the PageRank algorithm, we assume that all web pages have the same probability as the starting point or restart point of the random walk. In the Personalized PageRank algorithm, we assign different weights to each web page, enlarging or reducing their probability as a starting point, which is equivalent to giving them an initial importance. Corresponding to the anti-fraud scenario, we can then assign weights to

nodes (users or transactions, etc.) based on a small number of existing labels, thereby effectively utilizing the labeled data.

In a word, among the methods that combine graphs and machine learning, the most important thing is to extract features from graphs. CreditX, an Internet financial company based in Shanghai, summarized the process of combining machine learning and graph networks in the context of its business from the aspects of basic statistical characteristics and important features of graphs.

6.3.1.3 *GotCha — using PageRank method to extract global features*

In the social security system, which mandates that companies pay taxes to the government, literature found that some companies take advantage of loopholes of government systems to deliberately go bankrupt to avoid paying taxes.[87] The paper starts by linking companies to each other through shared resources to form a graph network and then goes on to find a spider web structure of fraudulent behavior.

This type of fraud refers to multiple fraudulent companies committing fraud under the control of a central company. The central company is unobservable and directly controls the resource allocation of other companies. The chapter uses a binary graph to construct this problem, as shown in Figure 6.11.

One of the most important steps in detecting fraud is to use the Personalized PageRank algorithm to extract global features. The global feature extraction method is as follows:

$$\overrightarrow{\xi_{k+1}} = \alpha \cdot Q_{\text{norm}} \cdot \overrightarrow{\xi_k} + (1-\alpha) \cdot \overrightarrow{z_{\text{norm}}} \tag{6.16}$$

where Q_{norm} denotes the expression of the normalized binary graph matrix. $\overrightarrow{\xi_{k+1}}$ denotes the global features obtained at the k+1st iteration. $\overrightarrow{z_{\text{norm}}}$ denotes the fraud label vector after normalization by node degree. α is a hyperparameter indicating the degree of individuality of the iterative process.

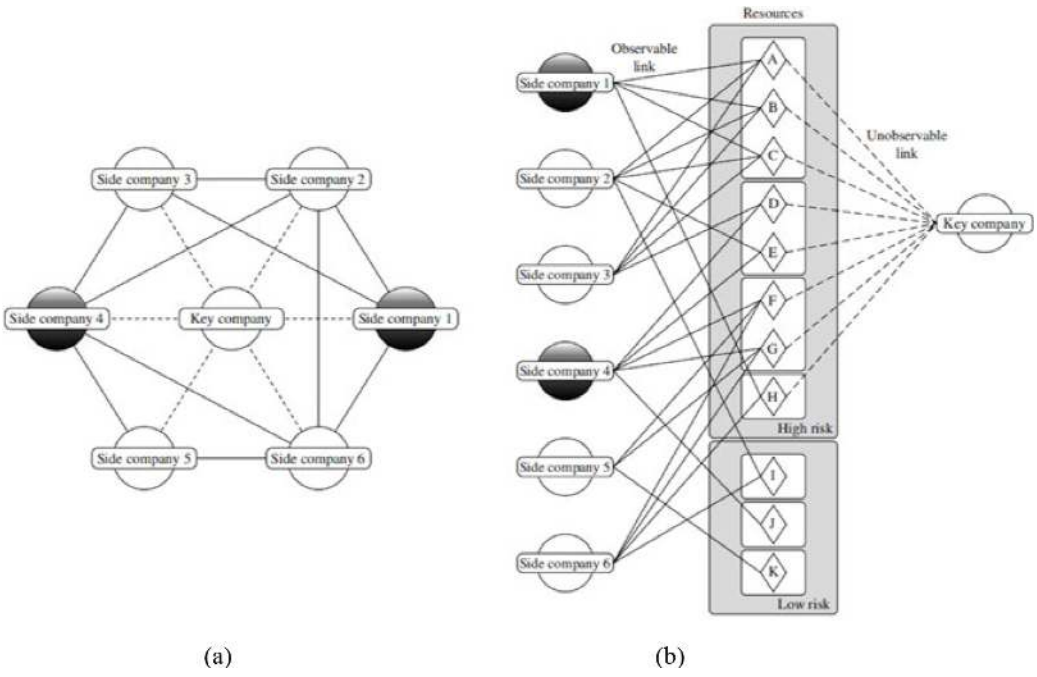


Figure 6.11. (a) Example of a spider web structure. (b) Bipartite graph of the spider web structure.

Figure 6.12 describes the computational flow of the GotCha method:

(1) Construct two forms of data of the node (company) at moment *t*: the node's intrinsic feature data and the graph dataset.
(2) Extract the features of two data structures: direct graph features (various statistical features), indirect graph features (global features obtained through the PageRank algorithm), and the node's (company's) intrinsic features.
(3) Combine the features and model the fraud problem using machine learning algorithms.

Three models are given in the paper:

(1) Intrinsic models containing only intrinsic features;
(2) A unipartite model that connects companies directly through resources that are shared or transferred between each other; and
(3) The bipartite model integrates company nodes and resource nodes to form a bipartite graph.

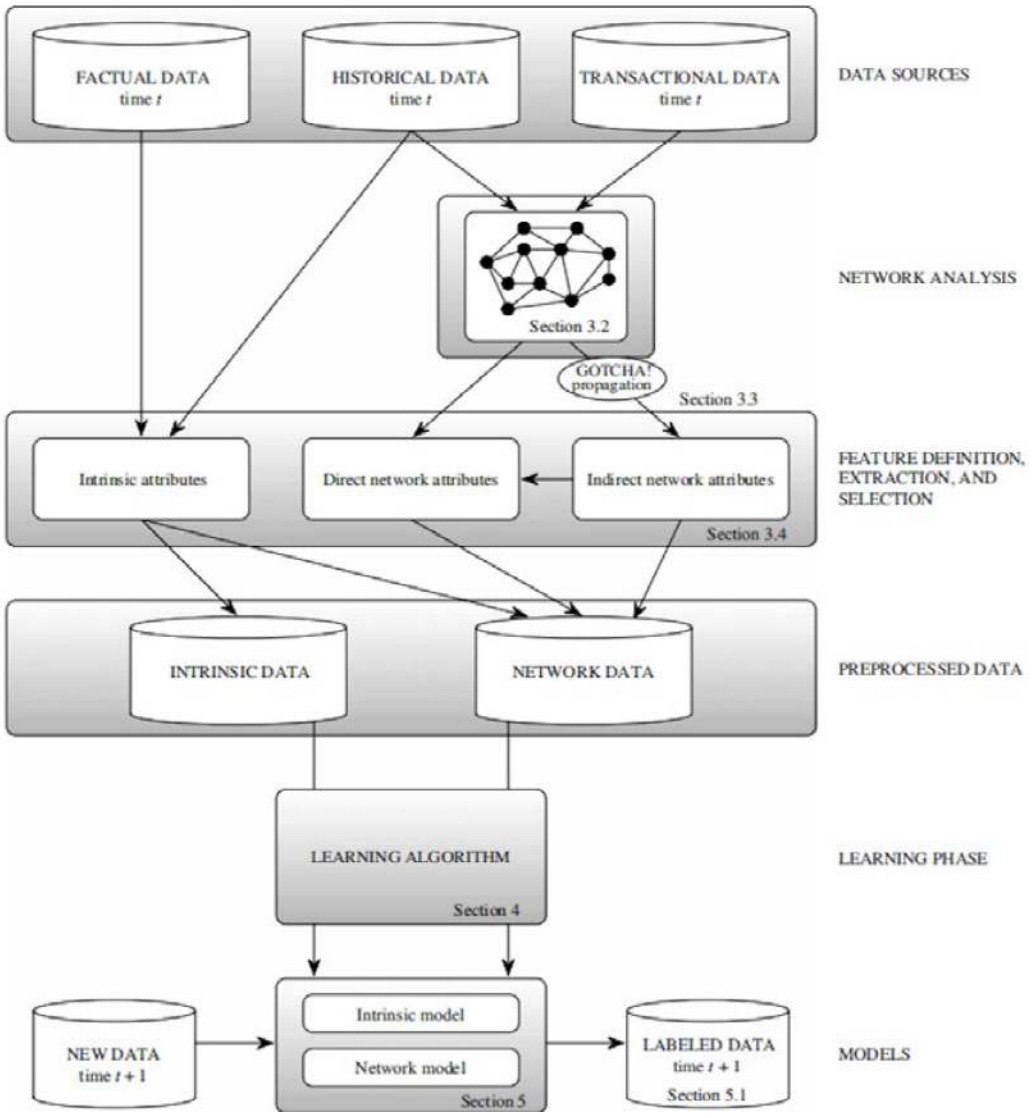


Figure 6.12. GotCha-model framework.

Figure 6.13 depicts the performance of the algorithm under different scenarios. It can be seen that the GotCha model effectively improves the testing capabilities of the model by incorporating global features.

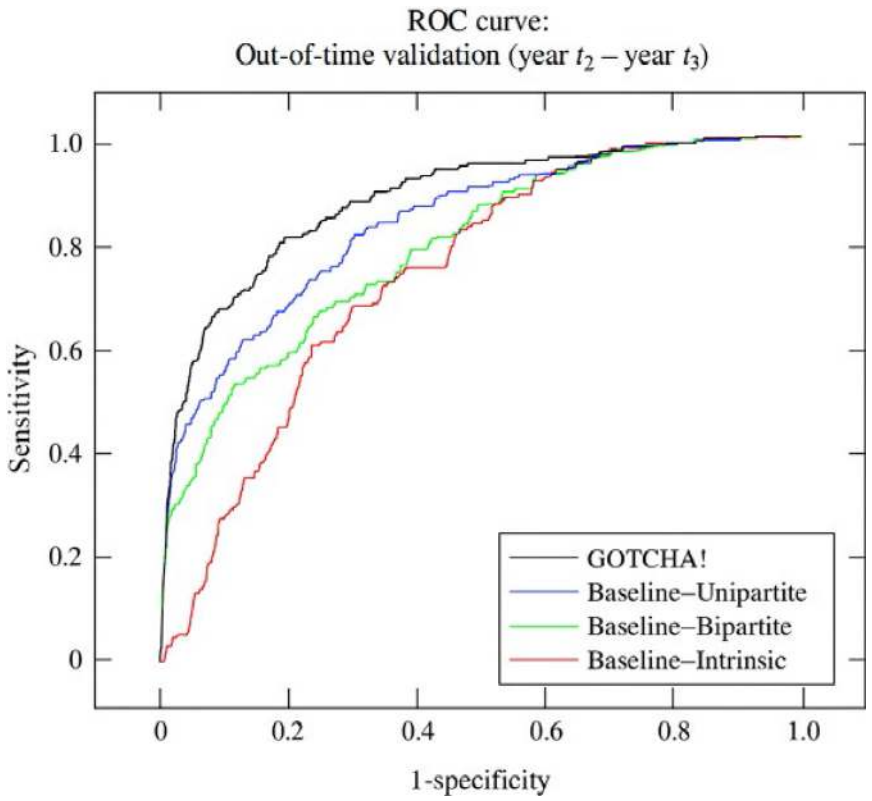


Figure 6.13. GotCha experiment results.

Note: All models are estimated on year t_2 and tested on year t_3.

6.3.1.4 *APATE — the first graph-based credit card transaction detection system*

The algorithm scenario in literature is fraud detection in the online payment process.[91] Fraudsters are able to learn spontaneously during the transaction process so that their transaction behavior becomes similar to that of ordinary non-fraudulent users. For online payment processing, a highly efficient and dynamic fraud detection system is necessary. Figure 6.14 describes the structural framework of the algorithm.

The paper proposes a feature extraction framework of Recency–Frequency–Monetary Value (RFM) framework for feature extraction based on the node's own features. Among them, "recency" is to use information of different lengths of time, "frequency" is mainly to measure the degree of dependence of different individuals on credit cards, and "currency" is mainly to distinguish the popularity of credit cards in different regions. In

short, the essence of this framework is to obtain different dimensions of features provided by different lengths of time by aggregating transactions over different time periods.

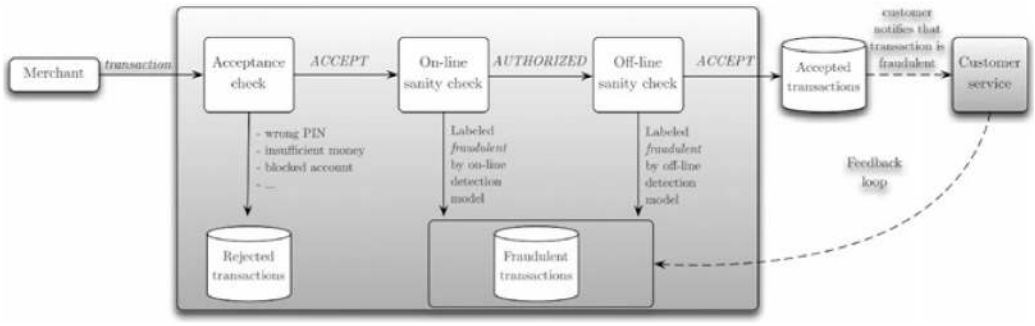


Figure 6.14. Model framework.

The network feature extraction method in the APATE system is similar to that in GotCha. The difference is that APATE also uses a tripartite graph in addition to using a bipartite graph to construct the matrix. Every transaction will involve both consumers and merchants. In addition to combining consumers and merchants to form a bipartite graph through transactions, the paper uses a tripartite graph when solving for global features through propagation to meet the propagation standard format of the PageRank algorithm. This is actually a bipartite graph that integrates merchants and consumers.

The specific process is as follows:

Firstly, we define a matrix that transforms the tripartite graph into a unipartite graph:

$$Q^{tri} = \begin{pmatrix} 0_{(c+m)\times(c+m)} & M \\ M' & 0_{t\times t} \end{pmatrix}$$

where the tripartite matrix $M_{(c+m)\times t}$ represents the matrix in which merchants and consumers are linked through transactions.

Secondly, the exposure scores of merchants, consumers and transactions are calculated through the PageRank propagation algorithm.

$$\overrightarrow{\xi}_k = \alpha \cdot Q_{\text{norm}}^{tri} \cdot \overrightarrow{\xi}_{k-1} + (1-\alpha) \cdot \overrightarrow{z}_{\text{norm}}^{tri} \tag{6.17}$$

Different global feature vectors are obtained by setting different α values.

When a consumer has not transacted with a merchant, we estimate the risk by using the exposure scores of its immediate neighborhood to calculate the exposure score of the transaction:

$$TXN_{i,k,\text{score}} = \frac{1}{\sum_{j=1}^{n} w_{i,j} + 1} CCH_{i,\text{score}} + \frac{1}{\sum_{j=1}^{m} w_{k,j} + 1} MC_{k,\text{score}} \tag{6.18}$$

where $W_{c \times m} = (w_{i,j})$ denotes the bipartite graph matrix consisting of transactions between merchants and consumers.

Finally, we estimate via a machine learning algorithm:

$$P\left(Y = fraud | X_{\text{Intrinsic}}, X_{\text{Network}}, X_{\text{Demographis}}\right)$$

When constructing bipartite and tripartite graphs, the values of their matrices represent values that decay over time. Furthermore, using a time-varying method to train the model effectively reduces the amount of data in the model. The specific method is to train the model only for n periods of time forward of the current time. The specific dynamic training process is shown in Figure 6.15, thus satisfying the need for a dynamic and efficient model as proposed in the paper.

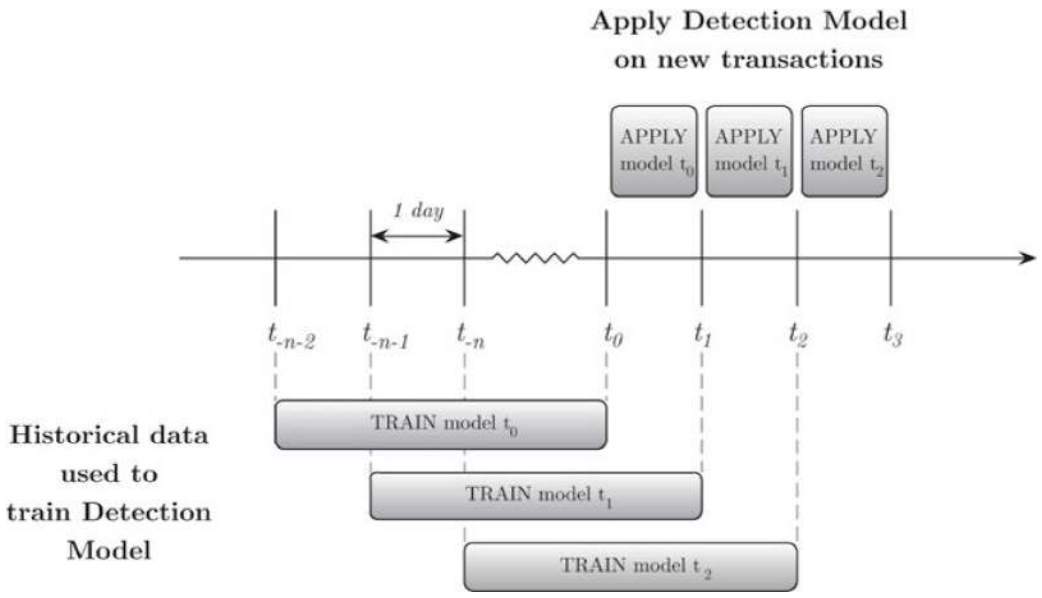


Figure 6.15. Time-varying training.

6.3.1.5 *CreditX — application of graph model in finance*

Literature provides an exhaustive description of feature extraction methods for graph models and introduces more data sources to build the network, including behavioral data, browsing data, and social data.[92] Based on this paper, we can sort out the basic methods of graphical modeling.

Traditional machine learning methods construct graphs through various data sources so as to obtain the graph features and individual features. These features are then used to build a machine learning model. Figure 6.16 illustrates the general modeling approach of combining graph models with machine learning algorithms.

Feature extraction for graphs can be categorized into three types: local features, global features, and expert features.

Local features can be described using degree correlation features, quadratic features, and local clustering coefficients of the graph. Global features can be extracted using methods in GotCha or APATE. Expert features are an extension to small data which can be implemented to improve the model. Table 6.1 shows the extraction of local features.

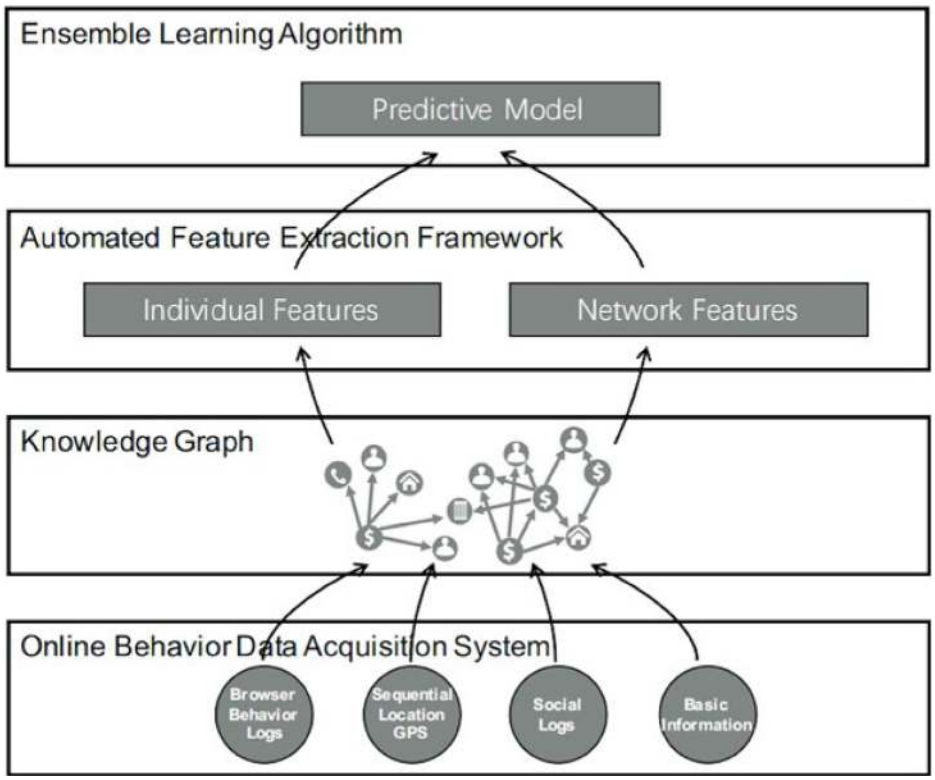


Figure 6.16. Combining graphical models and machine learning algorithms.

Table 6.1. Local feature extraction.

Metrics	Degree	Quadrangles	Local Cluster Coefficient
basic statistics	– number of associated applications	– total number of quadrangles – the max/mean/average of quadrangles frequency with the associated application nodes	– 2-order neighbors cluster coefficient – 3-order neighbors cluster coefficient
label dependent	– fraction of associated fraud applications	– fraction of quadrangles associated with fraud application nodes – the max/mean/average quadrangles frequency with the associated fraud application nodes	– 2-order neighbors (with only information nodes and fraud application nodes)cluster coefficient – 3-order neighbors (with only information nodes and fraud application nodes)cluster coefficient
intensity weighted	– weighted number of associated applications – weighted fraction of associated fraud applications	– weighted total number of quadrangles – the weighted max/mean/average quadrangles frequency of the associated application nodes – the weighted max/mean/average quadrangles frequency of the associated fraud application nodes	

The paper uses real-life data to conduct experiments and finds that models constructed from graphs can significantly improve detection capabilities.

6.3.1.6 *BotGraph — application of graph model in anti-"botnet" fraud*

The model in literature is designed to address an attack called "botnet," which is an attack in which a bot program can control many "broilers" and register a large number of email accounts through these "broilers" (bot-users/bot-accounts) and send spam.[93] We want to build algorithms that can find these bot-users.

Since the "broilers" are controlled by a small number of botnet commanders, they tend to have similar behavioral patterns and there will be resources shared among the bot-users. The proposed algorithm in the paper establishes the relationship between bot-users based on the IP addresses they use when registering, logging in, sending emails, etc. The main steps of the algorithm are as follows:

(1) Construction of User-User graph in which each bot-user is a vertex on the graph, and the number of IP addresses shared by two bot-users is the weight of the edge.
(2) **Top-down hierarchical clustering:** This involves setting up the edge weight (the number of shared IP addresses) and its threshold T. Then, the subgraph of G (edge weight $w \geq T$) is extracted, and graph clustering is performed on G to obtain k connected subgraphs. If the number of members of the connected graph is greater than M, further subgraph extraction is performed on the connected graph, and then the weight threshold will change from T to $T + 1$. This entire process is continuously iterated.
(3) **Pruning:** Once the group is obtained during the second step, evaluate the confidence level of the group, i.e., how suspicious the group is. The paper adopts a rule that considers a group a candidate bot-user group if 80% of its members send an average of >3 emails per day.
(4) **Grouping:** After the candidate suspect groups are identified in the third step, BotGraph traverses the tree again to solve the problem of which one should be selected if the parent node and the child node are both suspect groups. The tree traversal method is:

If the parent node (number of members is N) has at least one child node group containing no more than o(N) members, then traverse downward;

If the number of members of all child node groups of the parent node (number of members is N) does not exceed o(log N), then the traversing stops.

(5) **Verification:** This step aims to verify if the group has been correctly identified from multiple perspectives, such as ensuring that the account nicknames in the group follow a consistent pattern.

The algorithm above can be visually represented in Figure 6.17:

The top layer is the original User-User graph R, which is decomposed into two giant components (subgraphs), A and B in the second layer. The second layer also includes some other smaller subgroups or isolated vertices. The edges in components A and B all satisfy $w \geq T$, and the number of members is >M. Components A and B are then further decomposed into giant components C, D, and E. The edges of these subgroups all satisfy $w \geq T + 1$, and the number of members is >M. Components C, D, and E are further decomposed into small components (or isolated vertices) until there is no group that satisfies $w \geq T + 2$ and the number of members is >M.

Regarding the formation of components:

We recursively traverse from component A to its subcomponent C and stop traversing at component C as it satisfies the stopping condition. A consists of a giant component (C) and some other smaller components (or isolated vertices), and according to this method, A is a bot-user group (C belongs to group A).

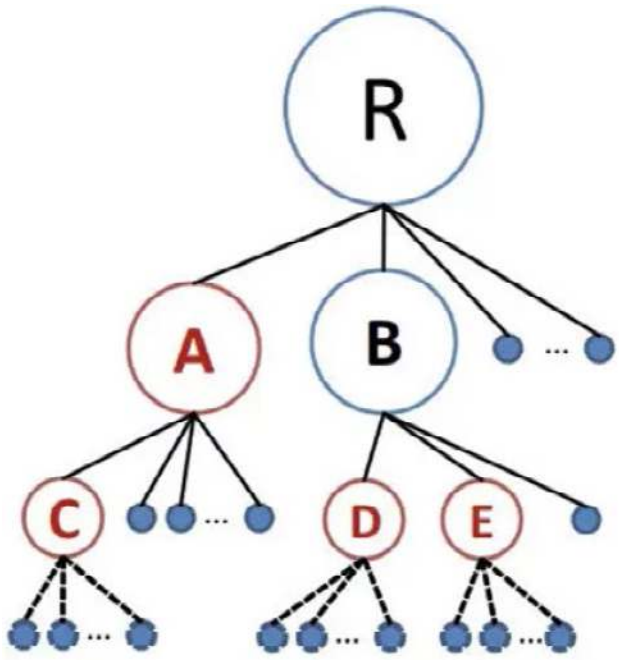


Figure 6.17. Algorithm process.

We recursively traverse from component B to its subcomponents D and E and stop traversing at components D and E as they both satisfy the stop traversing condition. B consists of two giant components (D and E) and some other smaller components (or isolated vertices), and according to this method, B is not a bot-user group, while D and E are bot-user groups.

6.3.2 *Applications of graph models combined with deep learning*

The application of graph model in deep learning methods can be roughly categorized into two types. The first type involves embedding the fraud recognition task into an existing network graph model algorithm. Applications of this kind can develop features that are more useful for the graph algorithm, thus enhancing the expressiveness of the graph model. Special embedding methods can also be utilized to improve the graph representation model. However, it is important to note that these methods must still adhere to the network graph model approach. Another approach is to choose the graph model based on the task at hand. During the task, the graph model can be selected for use in specific steps to achieve better

results. The algorithm framework as a whole must be established according to the task, with the network graph model being just one component.

These two methods are explained in more detail below. The literature is an example of the first method and involves adding features to the network graph model based on the task at hand.[94] Literature is an example of the second method and uses graph models in certain parts of the overall fraud detection framework to achieve improved results.[95]

6.3.2.1 *LongValue — Improvement and innovation of Node2Vec method*

Shanghai Dianrong Company has made some explorations into the Node2Vec algorithm for fraud identification in the P2P field in the literature, that is, using the Node2Vec method to automatically extract features from the representation method of the graph model.[90]

Their research uses the phone numbers in the address book to construct edges and generates the borrower's phone contact network; it uses the graph representation method to represent the nodes as vectors. The algorithm flow is as follows (Figure 6.18):

Original Node2Vec paths contain only node IDs, which means only the structure of the graph is considered. In fact, node attributes, such as third-party credit scores, gender, etc., are also important. They combine the properties into a LongValue, as shown in Figure 6.19. If the original attribute has many values, such as a credit score, it is divided into multiple levels. The combined attributes will be put into the model along with the vertex ID.

The paper uses *k*-neighbor algorithm and raw feature method to update the node vector, respectively and compare their performance with that of Node2Vec method on LongValue as showed in Figure 6.20. Although performance has been improved, the improvement is not significant.

In this regard, we analyze the reasons why the improvement is not significant:

(1) Too much noise — the phone numbers in the address book are generally very complex, and a large number of edges are basically useless. This makes it difficult to learn effective features through random walk. Thus, how to effectively filter noise becomes crucial.

(2) The data source is single, and it is not effective to build edges only through phone address books. Other sources, such as WeChat address books and public social media accounts, are also important sources. Multiple data sources enable filtering, making edge structures more effective.
(3) The Node2Vec method is useful for building features, however, it is not very useful for the fraud identification. Node2Vec is more inclined to learn the structural correlation characteristics of the network, and for fraud identification tasks pay more attention to the differences between the structure of fradulent nodes and that of other nodes. There are problems with the construction of local features and global features in the LongValue algorithm, which do not align with the task requirements.

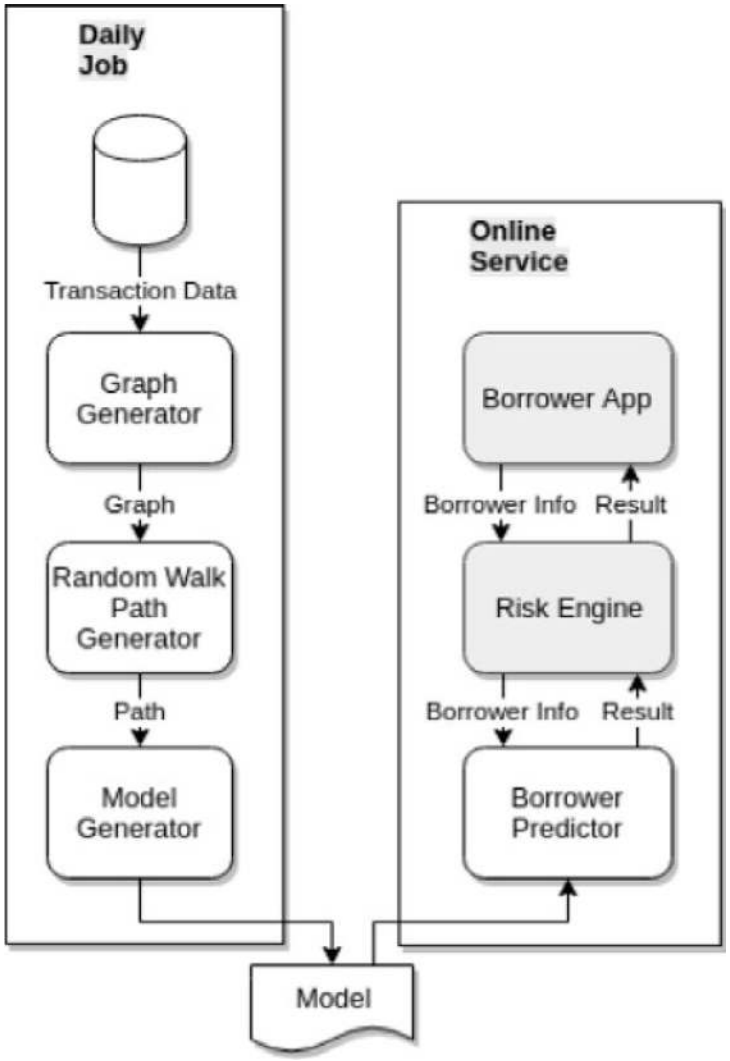


Figure 6.18. Algorithm process.



Figure 6.19. LongValue icon.

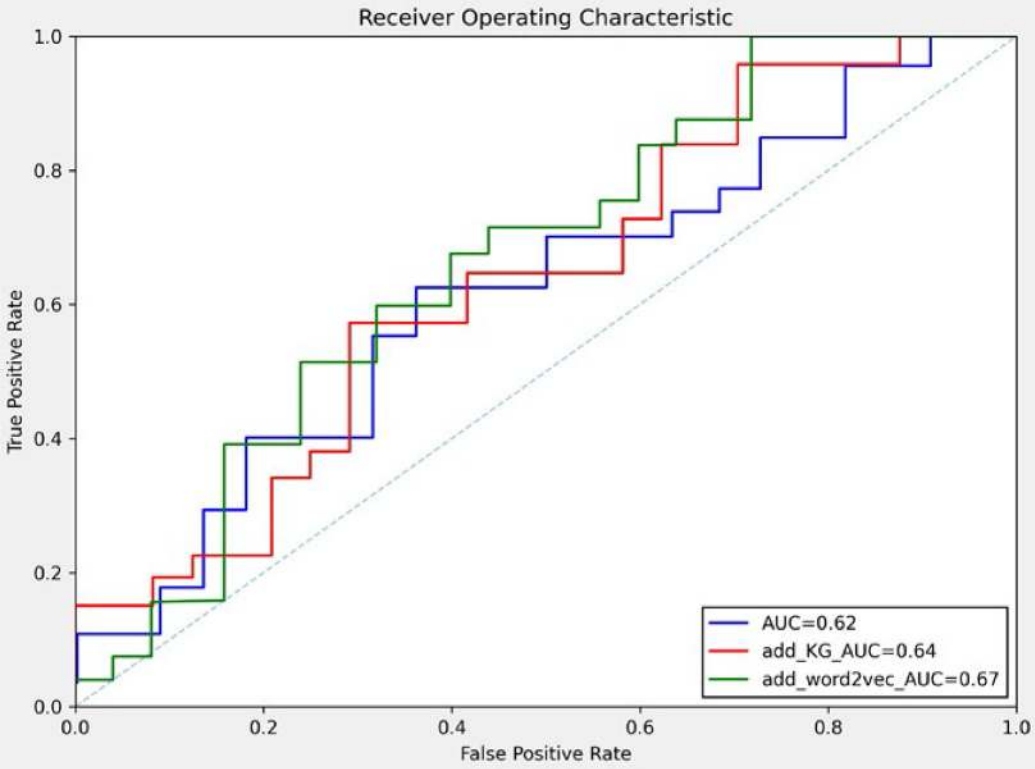


Figure 6.20. Effect comparison.

6.3.2.2 *FdGars — Fraud review detection in APP mall*

Tencent and Zhejiang University used graph convolutional networks to detect and explore false online reviews in literature.[94] Online review systems enable users to submit reviews about products; however, the openness of the Internet and monetary rewards for crowdsourcing tasks have encouraged a large number of fraudulent users to write fake reviews and post advertisements to interfere with the ranking of applications.

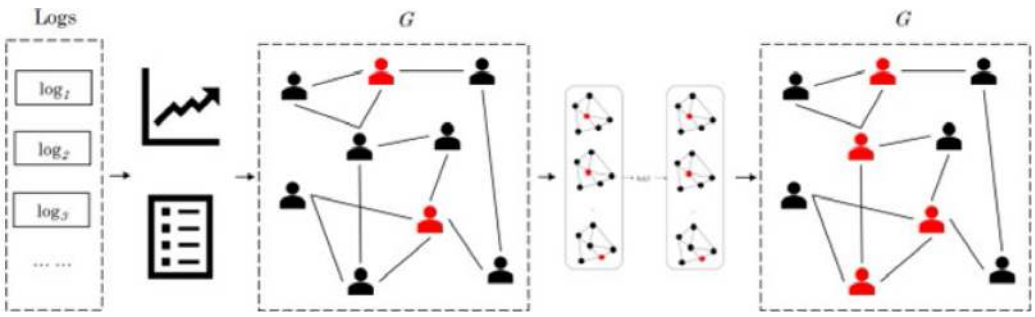


Figure 6.21. Algorithm framework.

The literature proposes a framework for identifying fraudulent reviews and compares the effectiveness of the algorithms. Its framework is shown in Figure 6.21.

The system first reviews the logs and extracts the review content and behavior of the reviewers; then establishes edges for reviewers who have reviewed the same APP within a certain time interval, thereby establishing the relationship between the reviewers; and finally builds the log into a graph Model G, $G = \{N, E, A\}$, where the three elements are nodes, edges, and the node's intrinsic characteristics.

The key to graph construction is identification. This chapter proposes a simple, rule-based fraudster identification method.

$$Label(n) = \begin{cases} 1 & \text{if } CD > \theta_{CD} \text{ and } DN > \theta_{DN} \\ 0 & \text{otherwise} \end{cases} \tag{6.19}$$

Among them, CD represents the number of consecutive days of posting comments within a period of time, and DN represents the number of devices used by the reviewer within a period of time.

This research uses a two-layer GCN structure, compares it with other algorithms, and demonstrates the effectiveness of the model. It conducts actual tests on Tencent's servers and finds that the application of graph convolutional networks is effective in Tencent's fraud identification framework. However, the research has obvious shortcomings:

(1) The identification method is too simple, and it is difficult to say that this simple identification is effective.

(2) The research lacks innovation and focuses simply on how to apply models.

6.3.2.3 *GraphRAD — Amazon's fraudulent account detection framework*

Amazon, in literature proposes a solution that combines artificial intelligence with system.[95] The system pushes possible fraudulent accounts to experts for judgment. According to this task model, the main task of the algorithm is to mine possible fraudulent communities. Since real accounts have characteristics such as excessively large correlation graphs and extremely unbalanced samples, GraphRAD's basic idea is to first reduce the size of the graph, then perform graph neural network calculations, and finally push potentially fraudulent accounts to experts for evaluation.

Their research propose GraphRAD — a graph-based risk account detection system. The system inputs a list of risky torrent accounts and outputs a short list of small, non-overlapping communities containing highly suspicious accounts. Its algorithm framework is shown in Figure 6.22.

The research offers three calculation modules:

(1) **Community detection module:** The community detection module uses the PageRank method to spread information about seed (fraud) nodes and detect communities through the conductance of the community;
(2) **Community extraction module:** This module merges all risky communities and grealy reduces the dimension of the graph. In order to reduce the overlap of the communities and reduce some fraud nodes that are not included in the community, the community extraction module uses the seed node propagation method to first cluster the merged graph, identify some non-seed nodes as seed nodes, and then Community detection is performed on the merged graph; and
(3) **Node scoring module:** There are still too many points that still remained after the two steps. Therefore, the paper uses the graph network method to score each extracted point and prioritize the high-risk ones for further inspection by human experts.

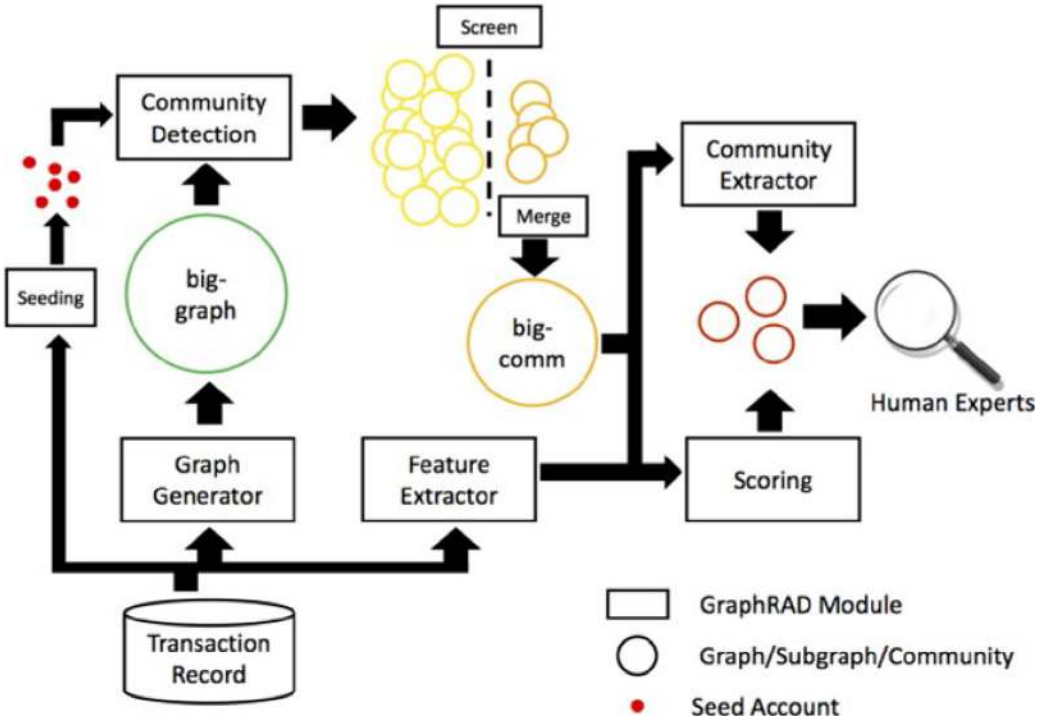


Figure 6.22. GraphRAD algorithm framework.

In the scoring module, three methods are used for evaluation. The evaluation results are shown in Figure 6.23. It can be seen that many graph-based learning models are almost unusable for large-scale graphs. For example, there are 60 million nodes and 500 million edges in this document. For the training of graph models, it's a huge challenge. Based on their own task settings, the researchers boldly delete a large number of nodes that are irrelevant to the task, reduce the amount of calculation, make the graph model more dynamic, and achieve the expected results.

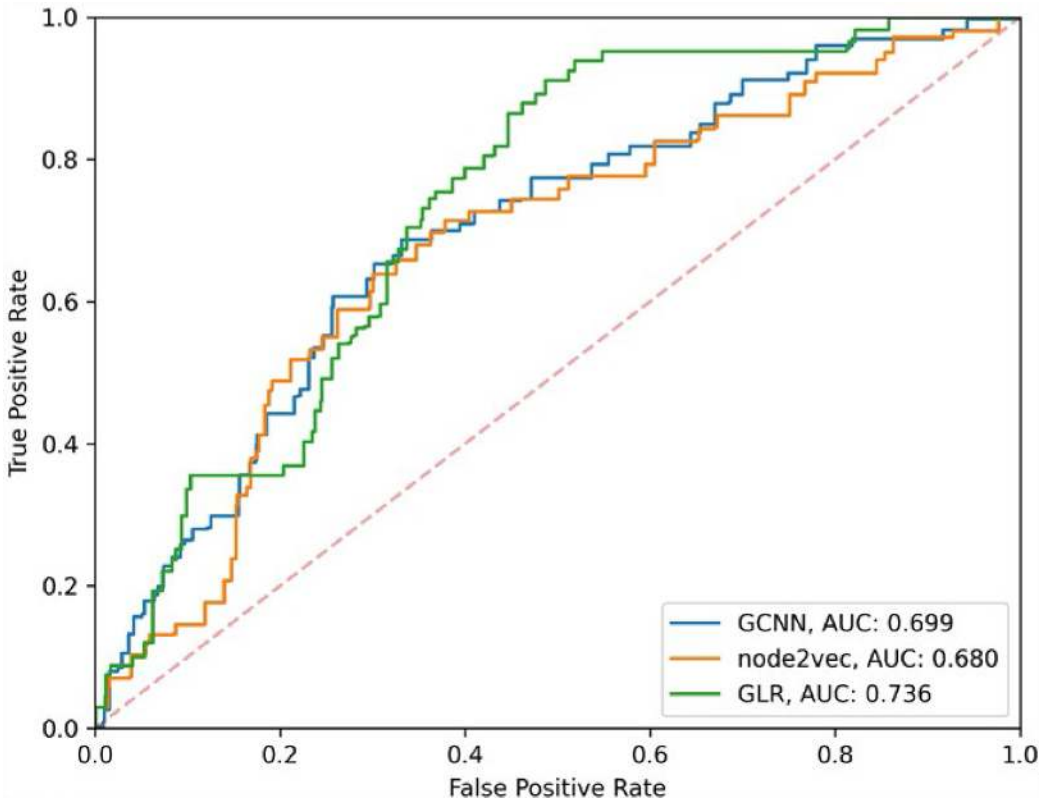


Figure 6.23. Evaluation result.

6.3.3 *Summary*

This chapter mainly introduces classic graph models and some of the latest graph neural network methods and their applications in the field of fraud identification.

In the node classification problem, the classic graph model can be divided into two types:

(1) The naive method is to first derive the adjacency matrix and merge its intrinsic features; then build a machine learning classifier based on the two features for classification.
(2) The collaborative method uses the idea of nodes propagating in the graph to propagate node information in the network; and then identifies probabilities for unidentified nodes.

Based on these two ideas, scholars have proposed different graph neural network methods to improve them.

The naive method considers the adjacency matrix to be a feature expression of the graph. In reality, adjacency matrices are often huge and sparse. If a machine learning classifier is used rashly, it will be difficult to train, and the model will fail. In order to avoid these problems, scholars have proposed various Embedding methods, that is, first using deep learning models to extract graph features, then using vectors to represent the graph features of nodes, and finally using machine learning classifiers for classification.

The collaborative method hopes to use the propagation characteristics of the graph to complete the classification of nodes during the propagation process. In reality, the combination of neural networks and graph computing requires huge computing resources. Therefore, after deep learning technologies such as convolution were recognized, scholars began to integrate the convolution modules in deep learning into graph neural networks. As a result, various efficient graph convolutional neural networks emerged.

However, these direct inference methods face computational constraints with extremely large amounts of data. The method that requires calculation of an entire graph for one gradient update obviously lacks applicability. Therefore, scholars have successively proposed new graph calculation methods, that is, partial calculation of the graph through the aggregation process and update process. This computing framework brings more possibilities to graph computing. For example, deep learning computing modules such as attention mechanism and LSTM were introduced.

In the community mining problem, this chapter introduces many cutting-edge algorithms, almost all of which can be seamlessly applied to the node classification problem. But in community mining, these algorithms are not so easy to use directly. We propose the following possible ways to apply the above algorithms to community mining.

For graphs, community mining can be formulated as a clustering problem, where clustering of nodes are carried out based on their similarity. For community mining, the representation vector of nodes is very important. Therefore, node embedding methods, which generated low-dimensional vector representations, can be effectively applied to community mining.

Another scenario is when some nodes are known to be fraudulent during community mining, and we need to mine the communities related to them. The label propogation method can be applied to address this problem, that is, to use node representation to simulate the propagation process on the graph. Alternatively, we can use a machine learning classifier to convert this mining problem into a semi-supervised learning problem, similar to node identification.

Chapter 7

Random Forest-Based Risk Identification of Internet Financial Subjects

7.1 Introduction

The rapid development of information technology has led to the extensive and in-depth application of cloud computing, social networks, and big data, resulting in the rapid rise of Internet finance as a new business format. The Internet financial industry, represented by peer-to-peer (P2P) online lending platforms, has unique advantages, and huge development potential, and has attracted a great deal of attention from government departments, industry, and academia.

P2P lending, also referred to as social lending or online lending, is the practice of lending money directly to individuals through online services. The business model involves establishing a third-party service platform to provide financing, information transmission, and related services for customers' investment and financial management needs. The borrower submits financing applications and related information to the platform operator through the P2P service platform. This includes mortgage information, loan amount, use of proceeds, loan term, credit report, cash flow, and repayment source. Investors provide information through the platform operator and decide whether to lend and if the funds are successfully matched, the platform operator will generally assist both parties in signing legal agreements and related mortgage procedures to pledge collateral.

Zopa, the first P2P online lending platform, was established in the UK in 2005. The Prosper platform was established in February 2006. It is the

world's largest P2P trading platform, with over 1.6 million members and funding of more than $500 million in loans. Lending Club is a US-based Internet loan company that is considered one of the most influential in the world. The P2P platform has since entered a stage of rapid development. CreditEase, established in China in 2006, was the first company to provide online loans in China, but its business is mainly offline. Paipaidai, established in 2007, conducted all transactions online from the beginning. Following that, the Chinese P2P industry experienced rapid development, with the highest number of loans and investors in the world. As of December 2016, there were an average of 2,448 operating platforms for loan businesses, and the P2P online loan market size reached 2,889.9 billion yuan.

According to statistics from the big data team at the Internet Finance Institute of Zhejiang University, 85% of the 2,189 P2P platforms across the country are concentrated in Guangdong, Beijing, Shandong, Zhejiang, and Shanghai. This high geographical concentration has resulted in relatively slow development in remote provinces in the west. Currently, 826 platforms are experiencing issues, accounting for a high proportion of 37.7%. Hainan, Shandong, and Guangxi are currently the regions most affected.

Online lending and small loans using P2P models are becoming more prevalent, providing convenience to investors but also posing significant risks.

On November 7, 2016, the State Council released a document emphasizing the need for stricter supervision of Internet finance. The government aims to improve P2P development, integrate Internet finance into the financial regulatory system, establish the legitimacy of its industry status, and regulate industry development to protect investor safety. Traditional banks have actively sought cooperation with standardized online lending platforms. However, as an Internet platform, the P2P financial platform is still in the development stage. Financial fraud incidents occur from time to time due to incomplete regulations and laws. To promote the healthy development of the P2P online lending industry, it is urgent to identify and mitigate risks.

7.2 Random Forest Method

The random forest method is one of the tree-based self-boosting methods first proposed in 2001. It is also one of the models commonly used in data science competitions. Its algorithm assumptions are weak; the model is effective with high flexibility and has a wide range of application prospects. The random forest method mainly integrates multiple decision trees through the Bagging method.

7.2.1 *Decision tree method*

Decision tree is a common machine learning method. Taking binary classification as an example, if you want to learn a model from a given dataset to classify new samples, then this classification task can be regarded as the problem of "whether the current sample is a positive example." The decision-making process is used to simulate people's decision-making methods. Usually, one dimension is selected for judgment. After the judgment is completed, other dimensions are selected. For example, if we want to make a judgment whether the borrower will be overdue, then we first check whether the person has passed the authentication. If this person has passed the authentication, then we check whether there is any loan history, etc. Generally, a decision tree contains a root node, several internal nodes, and several leaf nodes. Leaf nodes correspond to a decision-making process. The pseudocode is shown in Figure 7.1.

It can be seen from the pseudocode that the key to the decision tree lies in how to choose the optimal partition attributes. Generally speaking, we hope that the samples contained in the branch nodes of the decision tree belong to the same category as much as possible, i.e., the node purity is high.

Among them, there are roughly three methods for evaluating the purity of selection: information gain, gain rate, and Gini coefficient.

Information Gain: An indicator of node purity measured by information entropy. Assuming that the proportion of the kth category in the current sample set D is p_k (k = 1,2, …, n), then the information entropy of D is defined as

$$\mathrm{Ent}(D) = -\sum_{k=1}^{n} p_k \log_2 p_k \tag{7.1}$$

Assume that the discrete attribute a has V possible values $\{a^1, a^2, \ldots, a^V\}$. If a is used to divide the sample set D, V branch nodes will be generated, among which the vth branch node contains all samples in D with a value of a^v in attribute a, denoted as D^v. This way, we can deduce the "information gain":

Pseudocode

```
Training sample set D = {(x1,y1),(x2,y2),...,(xm,ym)}
Attribute set A = {a1, a2, ....., an}
TreeGenerate (D,A):
    Generate node
If all samples in D belong to the same category C:
    Mark node as a C-class leaf node
return
end if

If the attribute set A is empty or all attribute values of D are the same:
    Mark node as the most common class
return
end if
Select the best partition attribute a* from A
for a in a*:
    Generate a branch for node, where Dv represents a subset of samples in D with a* attribute value
of a
    If DV is empty:
continue:
else:
    TreeGenerate {Dv, A\ {a*}} recursively continues
end if
end for
```

Figure 7.1. Decision tree pseudocode.

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{D} \mathrm{Ent}(D^v) \tag{7.2}$$

The greater the information gain, the greater the "purity improvement" obtained by using a certain attribute for classification. The ID3 learning

algorithm in the decision tree uses information gain to select attributes for classification.

Gain Rate: The information gain metric is biased for attributes with a larger number of possible values. In order to minimize the possible adverse effects of this bias, the C4.5 decision tree algorithm uses the gain rate to select the optimal attributes for partitioning. The gain rate is defined as

$$\text{Gain}_{\text{ratio}(D,a)} = \frac{\text{Gain}(D,a)}{IV(a)} \tag{7.3}$$

where

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{D} \log_2 \frac{|D^v|}{D} \tag{7.4}$$

is called the intrinsic value of attribute *a*. The greater the number of possible values for attribute *a*, the larger the IV(*a*) value will usually be. However, the gain rate metric is biased for attributes with a smaller number of possible values, so instead of directly selecting the candidate partition attribute with the largest gain rate, C4.5 uses a heuristic: First, find the attributes with above-average information gain from the candidate attributes and then select the one with the highest gain rate among them.

Gini Index: The CART decision tree uses the “Gini index” to select partition attributes, where the Gini coefficient is

$$\text{Gini}(D) = 1 - \sum_{k=1}^{n} p_k^2 \tag{7.5}$$

The smaller the Gini index, the higher the purity of the dataset. Similarly, the Gini index of attribute *a* is defined as

$$\text{Gini}_{\text{index}(D,a)} = \sum_{v=1}^{V} \frac{|D^v|}{D} \text{Gini}(D^v) \tag{7.6}$$

In this chapter, the C4.5 method is used to build each tree in the forest.

7.2.2 *Ensemble learning method: Bagging*

Bagging is a method of self-sampling, a sampling with replacement conducted on the original samples. After multiple rounds of random samplings, we obtain a new sample derived from the sampling. This cycle is repeated, and finally, we can obtain T sample sets containing m samples. By building a base learning model on each sample set, we can then obtain T models. Finally, the T models are combined to predict the resulting output. The pseudocode is shown in Figure 7.2.

Pseudocode

```
The input is the sample set D={(x, y1), (x2, y2),... (xm, ym)},
Weak learner algorithm, weak classifier iteration count T. Output as the final strong classifier f (x)
for t=1,2.,T:
    Randomly sample the training set for the t-th time, collecting a total of m times
    Obtain a sampling set Dm containing m samples
end
For x in set Dm:
    Train the m-th weak learner Gm (x)
end
If classification algorithm prediction:
    The category or one of the categories with the most votes cast by T weak learners is the final category
If the regression algorithm predicts:
    The arithmetic mean of the regression results obtained from T weak learners is the final model value
output
```

Figure 7.2. Bagging pseudocode.

7.2.3 *Random forest*

Random forest is an extended variant of Bagging. On the basis of Bagging, which uses a decision tree as its base learner model, random forest further introduces the selection of random attributes in the training process of decision trees. Specifically, traditional decision trees will select all attributes when selecting partition attributes, while in random forests, for each node of each decision tree, an attribute subset of k nodes will be randomly selected. In the attribute subset, we choose the best attribute for

partitioning. The schematic diagram and pseudocode are shown in Figures 7.3 and 7.4, respectively.

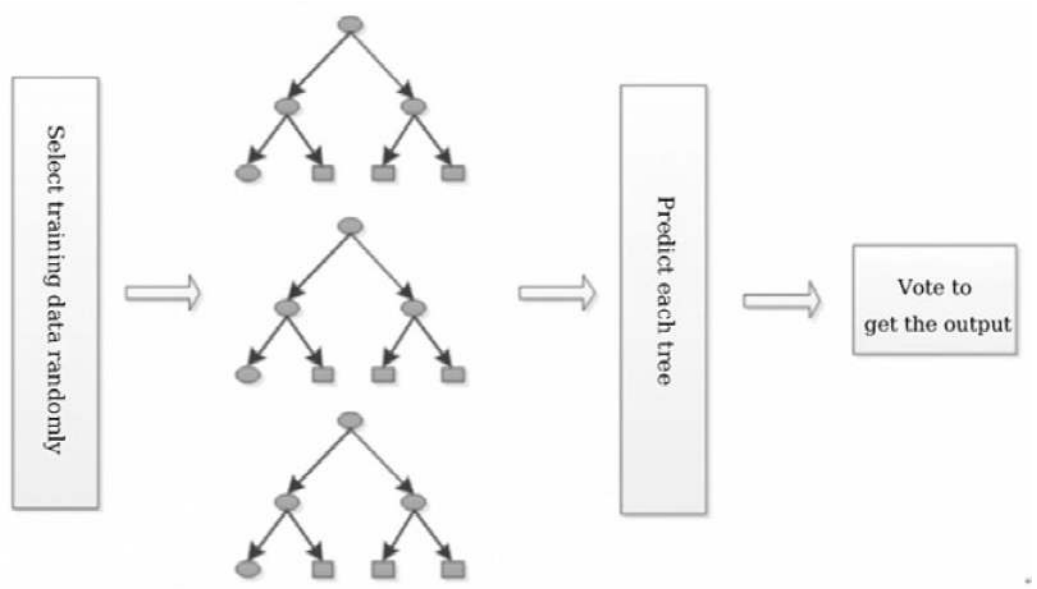


Figure 7.3. Summary of feature combinations.

Pseudocode

The input is sample set $D = \{(x1, y1), (x2, y2), \ldots, (xm, ym)\}$

Weak learner algorithm, weak classifier iteration count T. Output as the final strong classifier $f(x)$

for $t = 1,2, \ldots, T$:

    Randomly collect samples for the t-th time for a total number of m times

    Obtain a sampling set D_m including m samples

end

for x in the sampling size D_m:

    Use the sampling size D_m to train the mth decision tree model $G_m(x)$

    For node in all training nodes:

    Randomly select a subset of sample features from all the sample features on the node, and select the optimal feature from the randomly selected subset to generate the optimal node partition for making a decision tree

    Generate a $G_m(x)$

end

If classification algorithm prediction:

    The category with the most votes by T weak learners is the final category

If regression algorithm prediction:

    The arithmetic mean of the regression results obtained from T weak learners is the final model result

output

Figure 7.4. Random forest pseudocode.

7.3 Kernel Regression Method

The kernel regression method mainly maps data to high-dimensional space and solves problems in high-dimensional space. The kernel regression method is a non-parametric regression method that does not make any assumptions about the original data. By giving different weights to points with different distances from ∞, an intuitive loss function is obtained:

$$\text{Loss} = \sum\nolimits_i w_i(x)(y_i - c)^2 \tag{7.7}$$

From $\frac{\partial \text{Loss}}{\partial c} = 0$,

$$c^* = \frac{\sum_i x_i y_i}{\sum_j w_j} = \sum\nolimits_i \frac{w_i(x)}{\sum_j w_j(x)} y_i \tag{7.8}$$

$$c^* = \sum_i \frac{K(x, x_i)}{\sum_j K(x, x_j)} y_i \tag{7.9}$$

Commonly used kernels include the following:

(1) Polynomial Kernel:

$$K(x, z) = (x * z + 1)^p \tag{7.10}$$

Among them, p represents the polynomial order.

(2) Radial Basis Function (RBF) Kernel:

$$K\left(x, z\right) = e^{-\frac{\|x-z\|^2}{h}} \tag{7.11}$$

Among them, h represents the bandwidth.

(3) Laplacian Kernel:

$$K(x, z) = e^{-|x-z|} \tag{7.12}$$

The kernel regression method used in this chapter mainly uses RBF as the kernel, through which its standard deviation is computed in a high-dimensional space, with the aim of obtaining, through the kernel transformation, a value-at-risk based on each of the underlying subjects and thus applying the value at risk and expected return into the relevant financial theory.

7.4 Feature Combination Method Based on Frequent Patterns

7.4.1 *Frequent patterns and association rules*

Frequent patterns are itemsets, sequences, or substructures that occur frequently in a dataset. Frequent pattern mining is the basis of association rules, correlation analysis, and causal analysis and is also very helpful for classification and clustering. It is widely used in practice, such as market basket analysis, web log analysis, and DNA sequence analysis.

Every association rule is defined by two disjoint, non-empty sets X and Y. If there is $X \rightarrow Y$, it is said that $X \rightarrow Y$ is an association rule. For example, if you find that those who buy beer will also buy diapers, {beer}→{diapers} is an association rule. The strength of association rules is described by support and confidence:

Definition of Support:

$$\text{Support}(X \rightarrow Y) = \frac{\text{count}\left(|X \cap Y|\right)}{\text{count}(N)} \tag{7.13}$$

Definition of Confidence:

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{count}\left(|X \cap Y|\right)}{\text{count}(X)} \tag{7.14}$$

The higher the support and confidence, the stronger the rule. Association rule mining is to discover rules that meet a certain strength.

7.4.2 *Apriori method*

Existing relationships among discovered data can be defined by support and confidence. In the attribute list, there may be frequent itemsets for a single attribute, and there may also be frequent itemsets for two or more attribute groups. When calculating the support of a frequent itemset, it is usually necessary to traverse all product lists, which works well when the number of lists is small, but when the number of lists is tens of thousands, it becomes infeasible to traverse all lists because it would involve massive calculation. The Apriori method aims to solve this problem. The Apriori method is that if a certain itemset is frequent, then all its subsets must also be frequent. On the contrary, if an itemset is an infrequent itemset, then its

corresponding supersets are all infrequent itemsets. In this way, after it is determined that an itemset is a non-frequent itemset, the support of its corresponding superset does not need to be calculated anymore. This avoids the exponential growth of the number of itemsets to a large extent and allows for a more rapid calculation for mining frequent itemsets. The pseudocode is shown in Figure 7.5.

Pseudocode

```
Input: Dataset D, support threshold ∝
Output: Maximum k-itemset
Scan the entire dataset:
    Select itemsets with frequency k = 1
    If k = 0:
        Frequent itemsets are empty sets
While mining the frequent k-itemset:
    Scan data to calculate the support of each candidate k-itemset
    Remove the dataset with support below the threshold from the candidate k-itemset to obtain the
    k-itemset
    If the frequent k-itemset obtained is empty:
        Then directly return the set of k − 1-itemset as the result
        end
    If the frequent k-itemset obtained only has one item:
        Then directly return the set of k-itemset as the result
        end
    Generate k + 1-itemset
    Let k = k + 1
```

Figure 7.5. Apriori algorithm pseudocode.

7.4.3 *Apriori-based approach to feature combination*

When using algorithms for data mining, feature engineering is often involved. Feature engineering includes three parts: feature screening, feature combination, and feature construction. Feature screening methods are relatively mature with a variety of methods such as L1 regularization, statistical testing, and stepwise screening. The method of feature construction is relatively subjective and requires the creation of new features based on business-related background knowledge. There are different methods of feature construction in different fields and industries. However, currently, there have been no relatively standard and mature

feature combination processes. This chapter proposes a method of feature combination using Apriori. The specific process is to separate the two types of sample sets D_1 and D_2, and find the frequent pattern sets $Part_1$ and $Part_2$ that meet the required confidence and support levels; for $\forall count(r) > 2$, $r \in Part_1 \cup Part_2$, if $r \notin Part_1 \cap Part_2$, then r is the feature combination that meets the requirement. Finally, combine all the attributes in r as one feature. Only when all attributes in the combination appear, it will be identified as 1, and the rest will be identified as 0. The pseudocode is shown in Figure 7.6.

Pseudocode

Input: Dataset D
Output: Feature combinations that satisfy the frequency patterns
Divide the dataset into n categories, denoted as data groups D_n
Set support and confidence thresholds
For each data subset D_i:
    Generate its frequent pattern set P_i
    end
Compute the union of all P_i: P_a
Compute the intersection of all P_i: P_d
Compute the difference set: $P = P_a - P_d$
If the variable satisfies the elements in P:
    mark as 1
Else:
    mark as 0
Generate a new variable
output

Figure 7.6. Pseudocode of the feature combination method based on Apriori.

7.5 Fraud Identification Algorithm

7.5.1 *Data selection and data preprocessing*

There are numerous factors affecting the overdue rate of P2P lending, including borrower demographic and loan characteristics, macroeconomic environment, culture, and other factors. Based on a summary of the existing literature in academia, this chapter argues that there is a correlation between the borrower's demographic characteristics, borrower's historic payment

information, loan characteristics, platform authentication information, and P2P online lending overdue rate. This section mainly obtains P2P loan data using the crawler and the Paipaidai API. A total of 39 dimensions of data are obtained (Table 7.1), with a total data volume of 300,000 entries and a time span from December 2015 to February 2017.

According to Paipaidai's data, more than 95% of the loans that have been more than 30 days delinquent will default on the loan. Default here is defined as the failure to repay the loan by the end of the corresponding repayment period. Therefore, this section considers overdue payments of more than 30 days as fraudulent behavior in P2P lending.

Data Deletion: Since some of the data obtained are still within the repayment period, this chapter only selects data with due date before February 2017.

Data Labeling: If a loan is more than 30 days delinquent, it will be labeled as 1; otherwise, it will be labeled as 0. This way, the problem is converted into a binary classification problem.

Data Transformation: For categorical variables, computers can only recognize numbers, so we perform one-hot coding on categorical variables in this chapter and transform them into numbers that computers can recognize. For example, we convert {gender: male, female, unknown} into the form of {{1,0,0}, {0,1,0}, {0,0,1}}. The converted data can be used for calculations.

Training Set and Test Set Selection: Generally, the data are first partitioned, and the ratio of the training set to the test set is set to 3:1 for the data obtained. According to the degree of imbalance of the dataset, the Bagging method is used to divide the majority samples in the training set into 19 parts and then combine them with the minority samples to form 19 new balanced samples. We then use the 19 new samples to train the models separately and predict the samples in the test set. Finally, the prediction results of all models are averaged as the final output of the model.

Variable Mapping: The education variables appearing in the {undergraduate certificate, junior college, bachelor's degree, junior college

(vocational), master's degree, doctoral student, second bachelor's degree, high school diploma, evening college} are mapped to {1, 1, 2, 1, 2, 2, 1, 1, 1}, respectively. This mapping categorizes the education variables into low, medium, and high educational levels.

Table 7.1. Variable names.

Variable	Type	Description
FistBidTine	Date Time	Initial investment time
LastBidTime	Date Time	Last investment time
LenderCount	Int	Number of bidders
Auditing Time	Date Time	Date of transaction
Remainfunding	Decimal	Remaining investable amount
DeadLineTimeOrRemindTimeStr	String	Cut-offtime
CreditCode	String	Subject grade
Listingld	Int	List Number
Amount	Decimal	Loan Amount
Months	Int	Due date
CurrentRate	Double	Interest rates
BorrowerName	String	Borrower's user name
Gender	Int	1 male 2 female 0 unknown
EducationDegree	String	Education
GraduateSchool	String	Graduate school
StudyStyle	String	Format of study
Age	Int	Age
SuccessCount	Int	Number of successful bonowing
WasteCount	Int	Number of invalid bids
CancelCount	Int	Number of bid withdrawls
FailedCount	Int	Number of failed bids
NormalCount	Int	Number of normal payoffs
OverdueLessCount	Int	Number of overdue (1-15 days) settlements
OverdueMoreCount	Int	Number of overdue (more than 15 days) settlements
OwingPrincipal	Decimal	Principal remaining to be repaid
OwingAmount	Decimal	Amount to be repaid
AmmmtToReceive	Decimal	Amount to be received
FirstSuecessBorrowTime	Date	Time to first successful borrowing

	Time	
RegisterTime	Date Time	Registration Time
CertificateValidate	Int	Academic verification 0 Not Certified 1 Certified
NciicIdentityCheck	Int	Hukou verification 0 Not Certified 1 Certified
Phone Validate	Int	Mobile phone verification 0 Not Certified 1 Certified
Video Validate	Int	Video verification 0 Not Certified 1 Certified
Credit Validate	Int	Credit reference verification 0 Not Certified I Certified
Educate verification	Int	Education degree verification 0 Not Certified 1 Certified
LastSuccessBorrowTime	Date Time	Time to last successful bonowing
HighestPrincipal	Decimal	Maximum single borrowing amount
HighestDebt	Decimal	Highest debt in history
TotalPrincipal	Decimal	Cumulative borrowing amount

7.5.2 *Descriptive statistics*

In order to observe the relationship between categorical variables and fraud, this section calculates the percentage of fraud under different categories. Descriptive statistics of categorical variables were plotted in the bar charts in Figure 7.7. The following are some preliminary interesting conclusions drawn from the statistical graph in the figure:

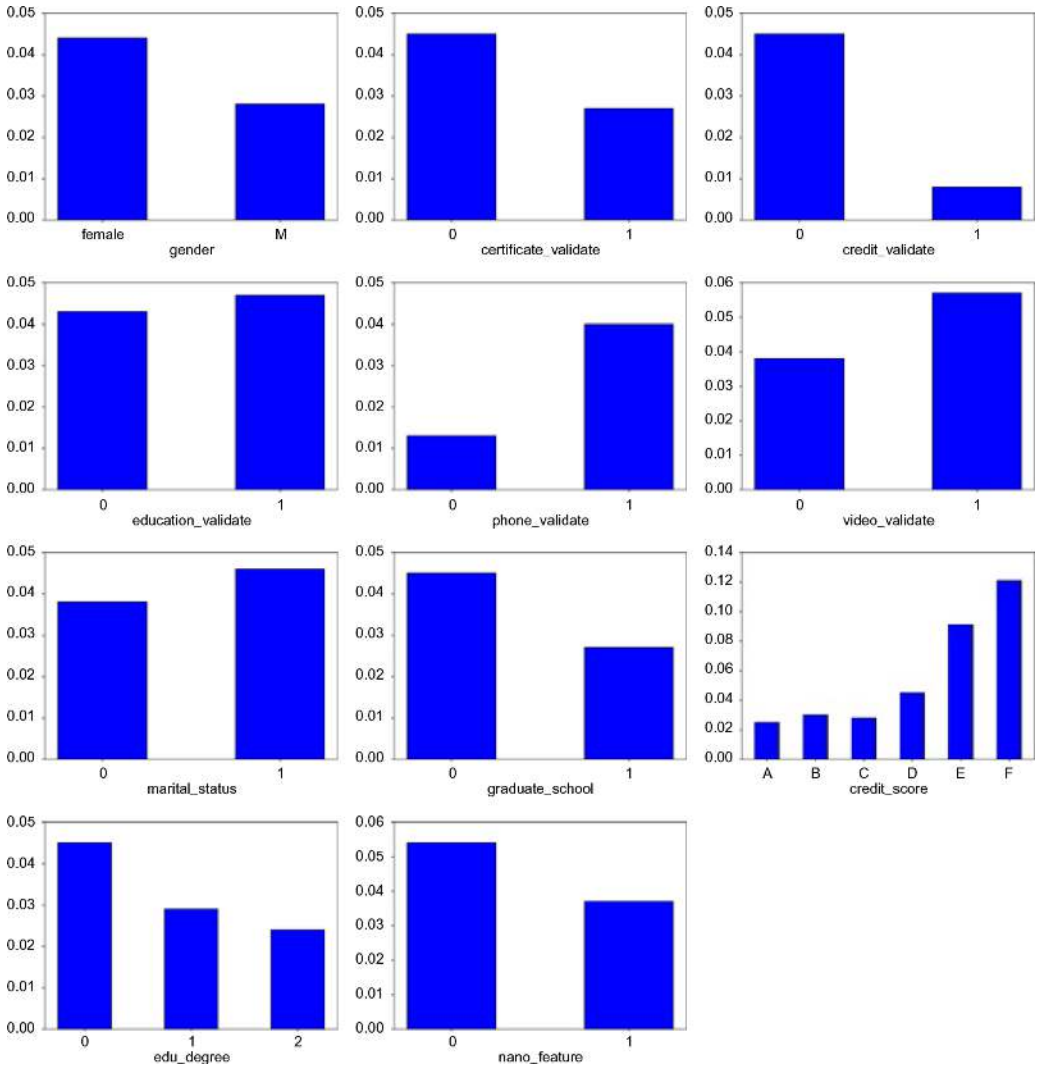


Figure 7.7. Category variable statistical chart.

(1) Women are less likely than men to commit fraud.
(2) Borrowers who underwent academic credential verification are significantly less likely to commit fraud than those who did not.
(3) Borrowers who underwent cell phone verification are more likely to commit fraud than those who did not.
(4) Paipaidai assigns ratings to loans and A is a higher rating than B and B is a higher rating than C. However, it seems that the fraud possibility of B-rated loans is higher than that of loans with C ratings and that the fraud possibility of C-rated loans is very close to that of A-rated loans.
(5) The higher the education level, the lower the likelihood of fraud.

(6) Borrowers who create their own usernames will be less likely to commit fraud than borrowers who use the random usernames assigned by the system.

The analysis of 11 categorical variables suggests that Paipaidai's remote authentication system needs improvement. Additionally, the risk ratings based on big data show bias in certain aspects. Academic credentials are also a differentiating factor. However, borrowers who passively verify their academic credentials by matching with information on the China Credentials Verification website using their identity have a higher probability of fraud than borrowers who actively verify their academic credentials, indicating that academic credentials alone are not as strong a differentiating factor as the active authentication behavior.

Figure 7.8 shows that almost all continuous variables have heavily skewed distributions. It is important to note that no conclusions can be drawn from a single distribution. To better understand the impact of continuous attributes on fraudulent behavior, the following section will use logistic regression to test it.

7.5.3 *Logistic regression statistical test*

As can be seen from Table 7.2, a total of 10 of the continuous variables are significant, of which the variables that have a positive effect on the probability of fraud are the amount to be repaid, the number of loan cancellations, the number of loan failures, and the history of being more than 15 days delinquent. It can be further illustrated as follows:

(1) The higher the outstanding balance, the higher the likelihood of fraud.
(2) If a borrower has a history of multiple loan cancellations, the higher the likelihood of fraud.
(3) The higher the number of loan cancellations, the higher the likelihood of fraud.
(4) The more instances of long-term delinquencies in history, the higher the likelihood of fraud.

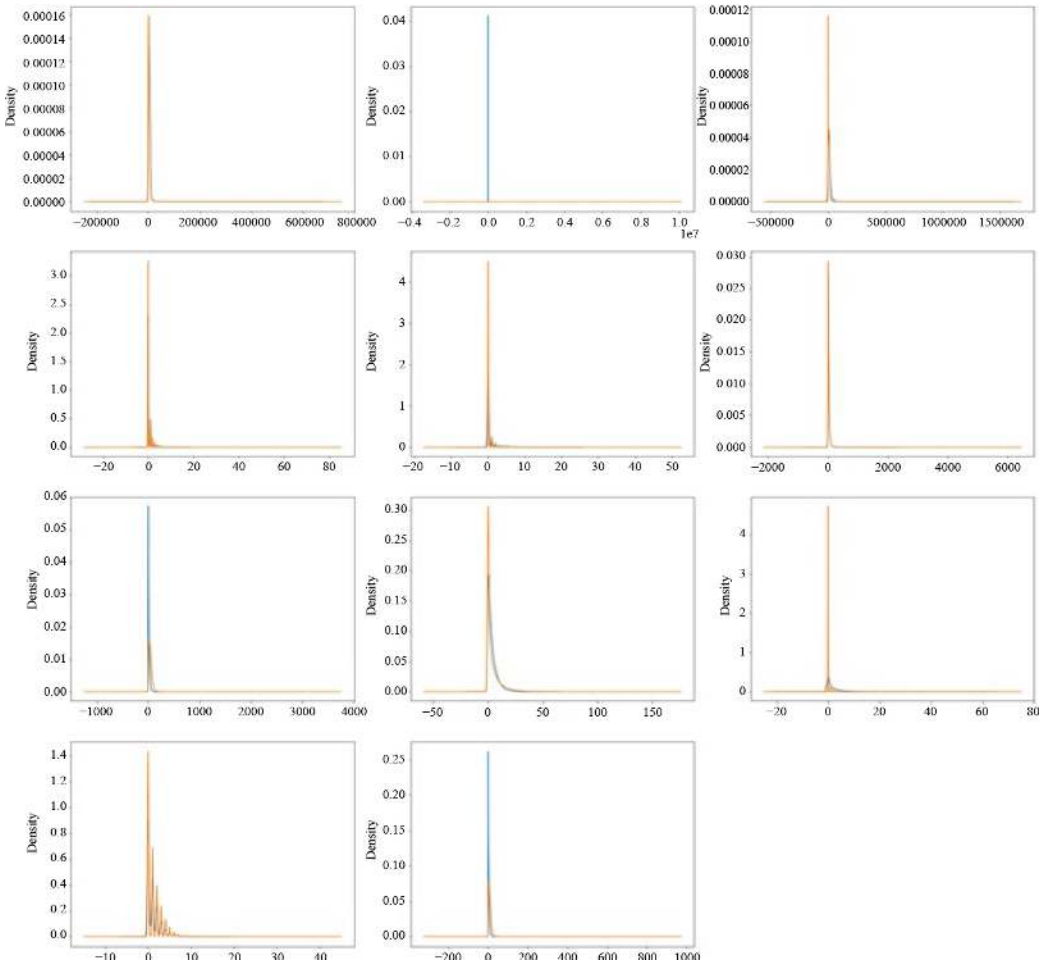


Figure 7.8. Distribution of continuous variables.

Table 7.2. Logistic regression statistical test.

	coef	std err	z	P value	Confidence interval
amout	−0.0003	0.0000	−54.4110	0.0000	[0.0000, 0.0000]
amounttoreceive	−0.0008	0.0000	−2.6740	0.0080	[−0.0010, 0.0000]
owingamount	0.0002	0.0000	79.3010	0.0000	[0.0000, 0.0000]
cancelcount	0.0734	0.0180	4.1850	0.0000	[0.0390, 0.1080]
failedcount	0.2191	0.0130	16.2750	0.0000	[0.1930, 0.2450]
lendercount	−0.0020	0.0010	−3.0670	0.0020	[−0.0030, −0.0010]
normalcount	−0.1289	0.0020	−60.6640	0.0000	[−0.1330, −0.1250]
overduelesscount	−0.0832	0.0040	−23.3400	0.0000	[−0.0900, −0.0760]
overduemorecount	0.5626	0.0070	78.0170	0.0000	[0.5480, 0.5770]
wastecount	−0.0068	0.0100	−0.7000	0.4840	[−0.0260, 0.0120]
successcount	−0.1522	0.0100	−15.4960	0.0000	[−0.1710, −0.1330]

Variables that have a negative effect on the likelihood of fraud are the amount borrowed, the number of bids pending, the number of bidders, the number of normal repayments, the number of short overdue payments, and the number of successful borrowings. This can be further illustrated as follows:

(1) The larger the original loan amount, the less likely the loan is to be overdue.
(2) The higher the number of bidders for the loan (indicating that the loan attracts more investors on the platform), the less likely it is to be overdue.
(3) If a borrower both invests and borrows money through the platform, the possibility of fraud is lower.
(4) The more regular repayments (which means the borrower has better credit), the lower the possibility of fraud.
(5) The more successful borrowings (which means the borrower has better credit), the lower the possibility of fraud.
(6) The less likely a borrower is to commit fraud if their payments are frequently overdue for a short period of time.

From the above analysis, some basic conclusions can be drawn in this section. Judging from historical borrowing behavior data and loan data, borrowers who are often overdue for a short period of time are less likely to commit fraud, which to a certain extent shows that Paipaidai has not done enough in late payment collection and has not reminded the borrowers of timely repayment. The low probability of fraud among borrowers who have successfully borrowed multiple times shows that borrowers with good credit have a certain degree of stickiness to the platform, indicating that the P2P method can attract a large number of borrowers with good credit rather than the online rumors that P2P is a gathering place for scammers.

7.5.4 *Algorithm parameter selection*

The identification of fraud is a binary classification problem on the model level. This section compares the effects of logistic regression versus random forest. First, the parameters of the two algorithms used are selected to achieve the best algorithm performance.

First, we select the penalty parameters of logistic regression. From Figure 7.9, we can see that when using the L1 regularization for parameter selection, when the penalty parameter is within a smaller range (in the 0.1–0.4 interval), there was a large fluctuation in performance accuracy and AUC of the algorithm, which indicates that some unimportant variables were not screened thoroughly enough when L1 yields sparse parameters. As the penalty parameter gradually becomes larger, the results of the algorithm tend to be stable, with the accuracy rate stabilizing at 0.91 and the AUC stabilizing around 0.9.

The random forest algorithm involves the selection of two important parameters. The first is the number of trees that make up the forest in the random forest, and the second is the depth of each tree. Theoretically speaking, random forest has a strong ability to prevent overfitting, so the more depth and number of trees the better, but the actual process will deviate from the theory. Furthermore, taking into account the computing resource consumption, the principle of parameter selection for the random forest algorithm in this chapter is to select a forest composed of the least number of trees and the shallowest depth of each tree for the best performance.

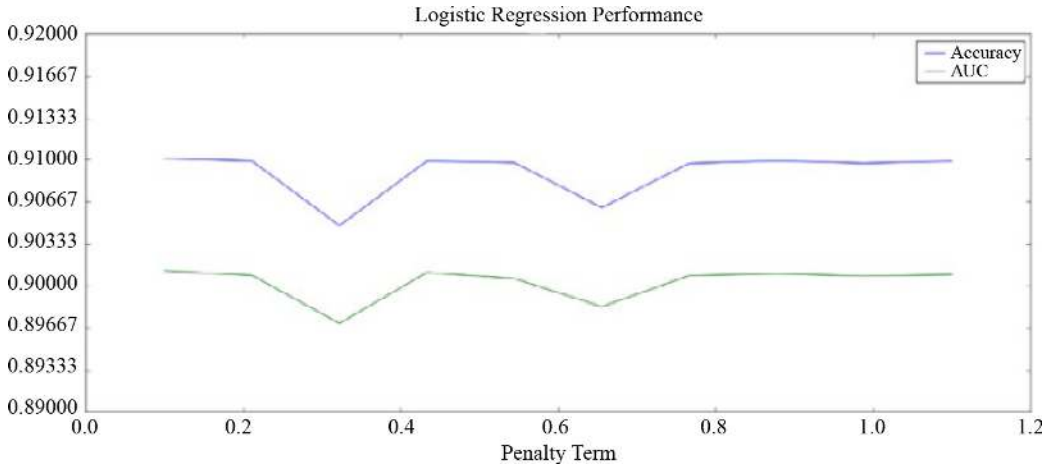


Figure 7.9. Performance of logistic regression algorithm.

Figure 7.10 depicts the performance of the algorithm for forests composed of different numbers of trees. It can be seen from the figure that the performance of the algorithm on the test set becomes better as the

number of trees increases. When it reaches 35 trees, AUC starts to have a relatively stable performance. When it reaches 60 trees, the algorithm gets an optimal performance on the test set.

Figure 7.11 depicts the performance of the random forest algorithm composed of trees of different depths. As seen in the figure, both the accuracy and AUC of the algorithm on the test set initially increase as the depth of the tree increases. Subsequently, AUC levels off after reaching 20 layers of depth, and the accuracy levels off after 25 layers of depth, so 25 layers of depth can be regarded as an optimal performance. The final algorithm keeps the accuracy around 0.921 and the AUC around 0.905.

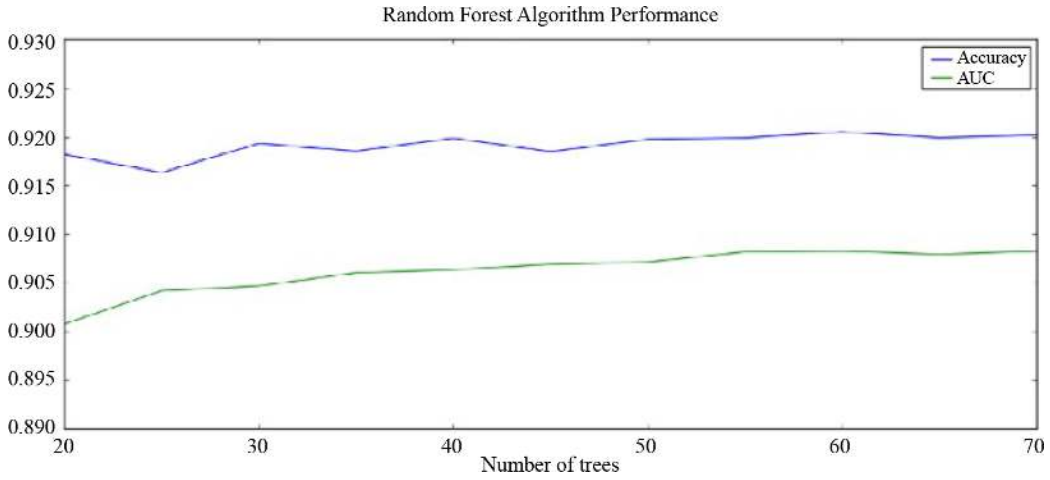


Figure 7.10. Performance of random forest algorithm.

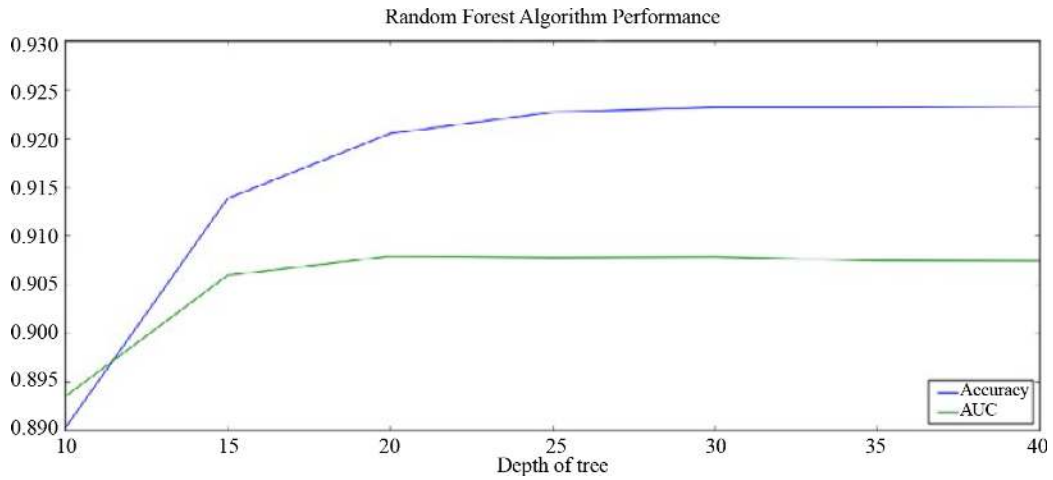


Figure 7.11. Performance of random forest algorithm.

Based on the above analysis, when comparing model effects in this chapter, 1 is used as the penalty parameter of logistic regression. In the random forest algorithm, 60 trees with a depth of 25 layers are used to form a random forest.

7.5.5 *Feature combination*

After completing the parameter selection, this section uses the Apriori algorithm to improve the algorithm on the categorical variables of the original data. By setting the support and confidence parameters, we search for frequent patterns for fraud and non-fraud categories, respectively, and compare different fraud categories. A total of 21 frequent patterns are found in this chapter, as shown in Table 7.3.

Setting the frequent patterns as a new variable would result in 21 more attributes in the original dataset. As can be seen from the table, all frequent patterns include education information, echoing the previous descriptive statistical analysis, suggesting that education is an important indicator of the borrower's likelihood of fraud. After improving the algorithm through feature combination, the accuracy and AUC in 19 datasets are significantly higher than the algorithm without feature combination, for which it can be inferred that the algorithm with feature combination performs better than the algorithm without feature combination. The results are shown in Figures 7.12 and 7.13.

7.5.6 *Feature screening and importance analysis*

The random forest algorithm selects nodes through random sampling. Redundant variables decrease the likelihood of selecting important variables, impacting the algorithm. Therefore, feature screening is crucial. Based on the algorithm described above, this section further performs feature screening based on variable importance and filters out two variables with low importance each time. The importance score in the random forest algorithm reflects the purity of the split of variables used by each decision tree from the leaf nodes, indicating their degree of usage. The higher the importance, the greater the impact of the variable on the results.

Figure 7.14 shows that the algorithm's accuracy and AUC improve after the variable screening begins. However, after the 23rd screening, both metrics begin to decline significantly. Therefore, it can be inferred that the variables at the 19th screening produce the best algorithm performance. Table 7.4 illustrates the importance of these variables.

Table 7.3. Summary of combined features.

Portfolio Name	Variable combination results
cond0	creditvalidate_0,graduateschool_0,nciicidentitycheck_0,phonevalidate_1,educatevalidate_0
condl	edu_degree_0,creditvalidate_0,nciicidentitycheck_0,phonevalidate_1,educatevalidate_0
cond2	videovalidate_0,creditvalidate_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond3	edu_degree_0,videovalidate_0,creditvalidate_0,phonevalidate_1,educatevalidate_0
cond4	videovalidate_0,creditvalidate_0,graduateschool_0,phonevalidate_1,educatevalidate_0
cond5	creditvalidate_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond6	edu_degree_0,videovalidate_0,creditvalidate_0,nciicidentitycheck_0,phonevalidate_1,educatevalidate_0
cond7	videovalidate_0,creditvalidate_0,graduateschool_0,nciicidentitycheck_0,phonevalidate_1,educatevalidate_0
cond8	edu_degree_0,videovalidate_0,creditvalidate_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond9	edu_degree_0,videovalidate_0,creditvalidate_0,graduateschool_0,phonevalidate_1,educatevalidate_0
cond10	edu_degree_0,creditvalidate_0,graduateschool_0,nciicidentitycheck_0,phonevalidate_1,educatevalidate_0
cond11	videovalidate_0,creditvalidate_0,graduateschool_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond12	edu_degree_O,creditvalidate_O,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond13	videovalidate_0,creditvalidate_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond14	creditvalidate_0,graduateschool_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond15	edu_degree_0,videovalidate_0,creditvalidate_0,graduateschool_0,nciicidentitycheck 0,phonevalidate 1,educatevalidate 0
cond16	edu_degree_0,videovalidate_0,creditvalidate_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond17	edu_degree_0,videovalidate_0,creditvalidate_0,graduateschool_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond18	videovalidate_0,creditvalidate_0,graduateschool_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond19	edu_degree_0,creditvalidate_0,graduateschool_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0
cond20	edu_degree_0,videovalidate_0,creditvalidate_0,graduateschool_0,nciicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0

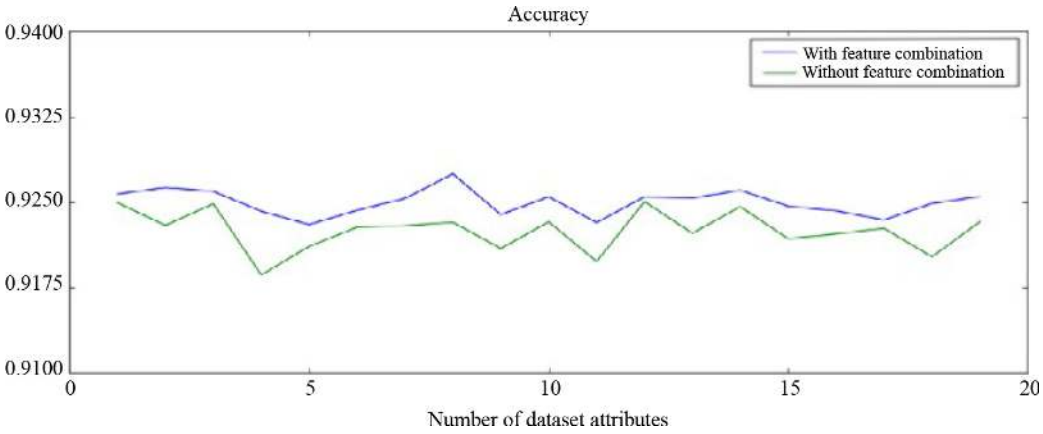


Figure 7.12. Accuracy performance before and after feature combination.

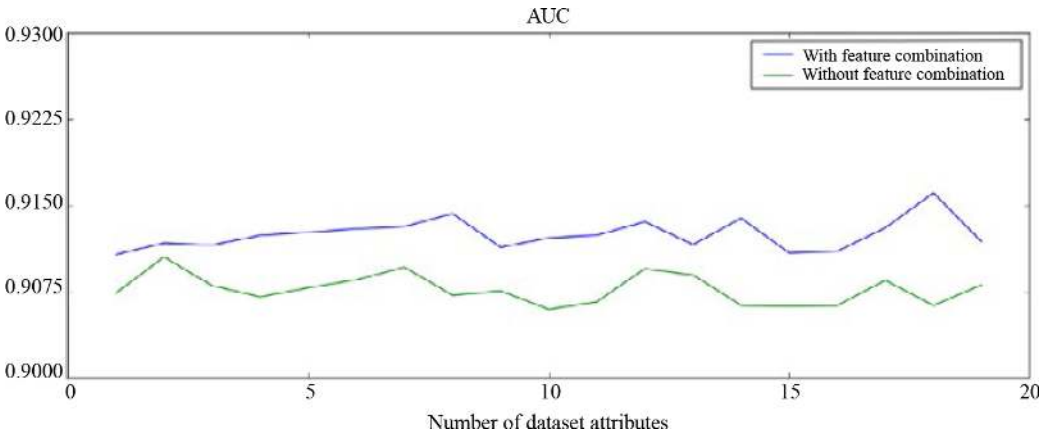


Figure 7.13. AUC performance before and after feature combination.

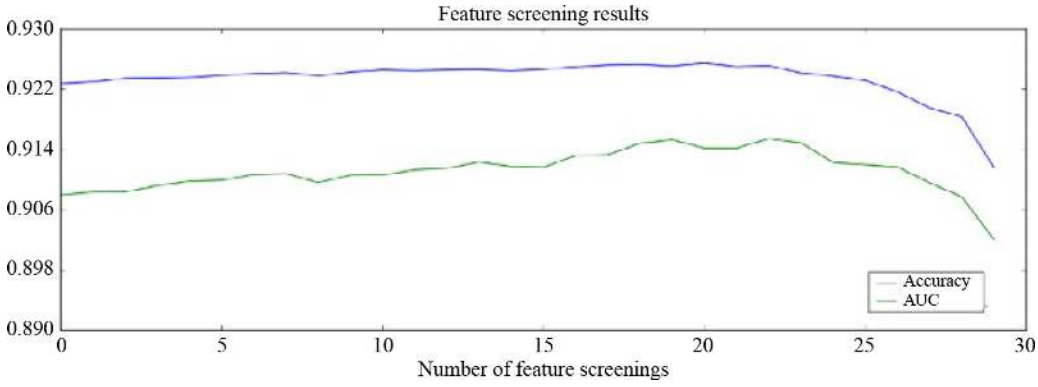


Figure 7.14. Feature screening results.

The six most important variables for assessing creditworthiness are the number of historical normal repayments, the number of historical long-term delinquencies, the unpaid principal balance, the number of historical successful borrowings, the number of historical short-term delinquencies, and the original loan amount. The results are generally consistent with established logic of risk assessment. The borrower's creditworthiness and stickiness to P2P lending can be inferred from the number of historical normal repayments and successful borrowings. Additionally, the borrower's ability to incur debt can be measured by the unpaid principal balance and initial loan amount. The number of historical long-term and short-term delinquencies are also variables that indicate the borrower's repayment behavior and ability.

Table 7.4. Importance of variables.

variable name	significance
normalcount	0.235064
overduemorecount	0.170422
owingamount	0.152772
successcount	0.116163
amout	0.057889
overduelesscount	0.045173
lendercount	0.038754
age	0.035946
wastecount	0.01991
currentrate	0.015079
failedcount	0.014982
creditvalidate_1	0.012852
cancelcount	0.010929
months	0.010341
creditvalidate_0	0.009823
gender_2	0.008522
cond8	0.007678
certificatevalidate_0	0.007583
cond11	0.006671
name_feature_0	0.006319
creditcode_D	0.005987
creditcode_C	0.005724

name_feature_1	0.005417

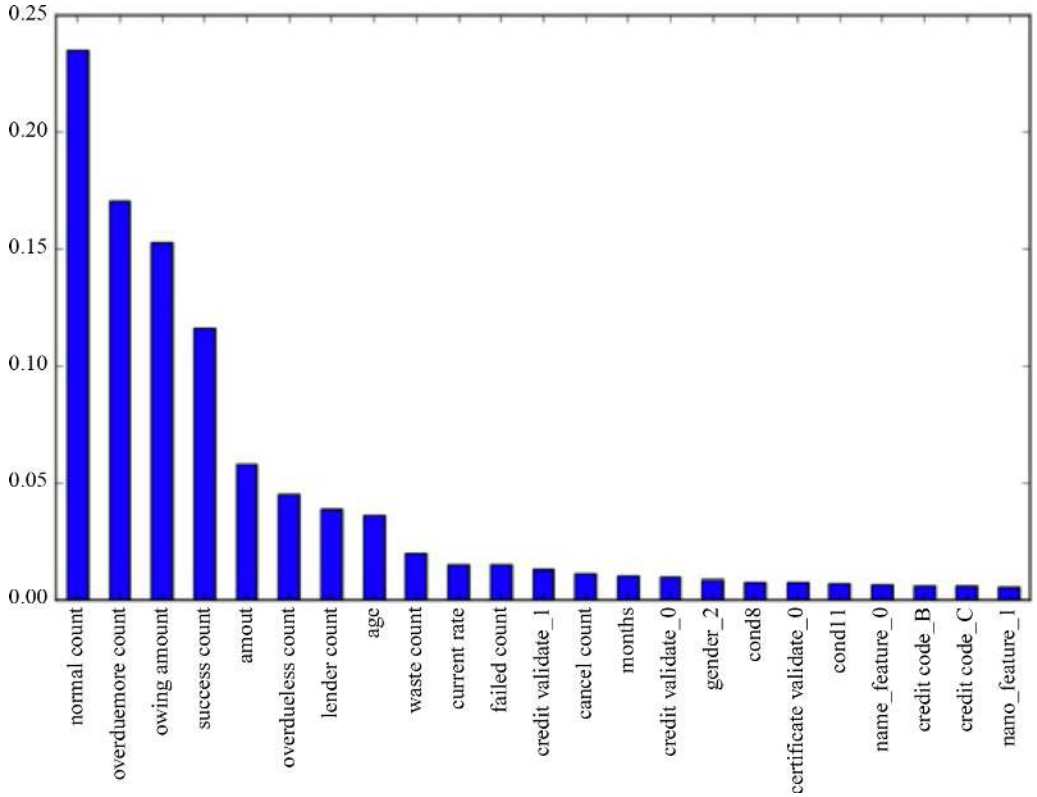


Figure 7.15. Importance of variables.

The education factor from the previous analysis does not appear in the importance analysis results, but according to Figure 7.15, we can see that there are two combinations of features that are of certain importance in measuring borrowing behavior:

Composition pattern of cond8: edu_degree_0, videovalidate_0, creditvalidate_0, certificatevalidate_0, phonevalidate_1, educatevalidate_0.

Composition pattern of cond11: videovalidate_0, creditvalidate_0, graduateschool_0, certificatevalidate_0, phonevalidate_1, educatevalidate_0.

It can be seen that the difference between the two combinations and the remaining combinations lies in the additional education levels and

graduation institutions. Therefore, the absence of education variables in the importance score is because education information has been included in the combined features, so the importance analysis results actually are not contradictory to the previous analysis. Furthermore, the possible reason why the education characteristics are not represented in the importance score is that education variables are very similar to other variables and can be represented by other variables. Since random forest is a non-linear algorithm, its expression method may be non-linear. The new combinations of features are not replaced in the random forest algorithm, thus supporting the effectiveness of the feature combinations in this algorithm.

7.5.7 *Algorithm comparison*

In order to select the final risk identification model, this chapter compares the two models (Figures 7.16 and 7.17). Judging from the performance of their accuracy and AUC, the accuracy of random forest reaches around 0.926, while the accuracy of logistic regression is around 0.91. The AUC of random forest reaches around 0.913, while the AUC of logistic regression remains around 0.90. Obviously, random forest is significantly better than logistic regression in fraud detection. Therefore, we chose random forest as the fraud detection algorithm.

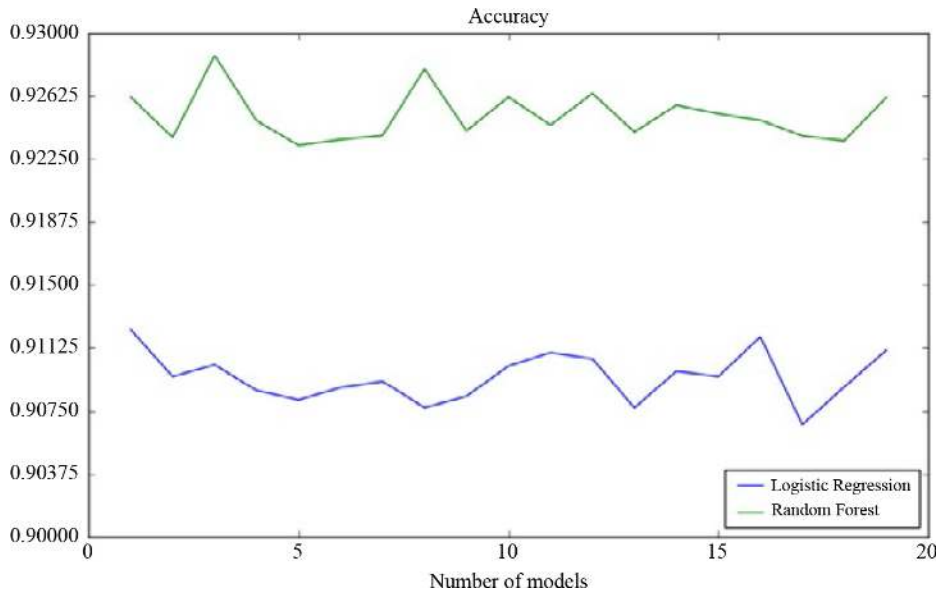


Figure 7.16. Comparison of accuracy between logistic regression and random forest.

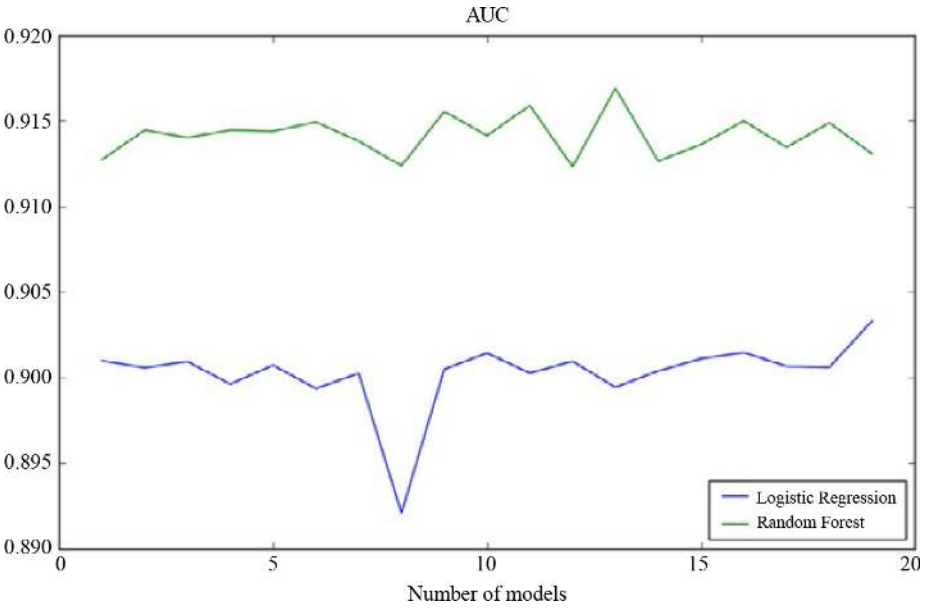


Figure 7.17. Comparison of AUC between logistic regression and random forest.

Table 7.5. Confusion matrix.

	Predicted nonfraud	Predicted fraud
Actual nonfraud	62148	5045
Actual fraud	280	2459

The confusion matrix of the test set (Table 7.5) shows that the random forest based fraud detection algorithm correctly identifies approximately 90% of the actual fraudulent loans, demonstrating a high level of accuracy.

The above research on P2P risk identification and corresponding investment models utilized crawler technology and OpenAPI to obtain the latest real P2P data, reflecting the actual Chinese investment environment. Machine learning algorithms were used to study these data, revealing that academic credentials, borrowing history data, and borrowing behavior significantly impact fraud. A fusion algorithm that combines the Apriori algorithm and the random forest algorithm is proposed. Empirical analysis shows that this new algorithm significantly outperforms the logistic regression model.

7.6 Investment Recommendation Algorithm

7.6.1 *Investment recommendation algorithm based on kernel regression*

The algorithm first observes the probability distribution of fraud. Figure 7.18 shows that the distribution is severely skewed, with most of it concentrated in the range of 0–0.2, and the overall fraud rate is controlled at around 4%. This indicates that the fraud rate in the Internet P2P industry is not very high. Overall, Paipaidai has good risk control. Considering that although the algorithm's identification of fraudulent loans is accurate, there may be a potential for bias when the algorithm is used for investment. The kernel regression method is utilized to determine the potential deviation risk of a loan. Assuming that the calculated value from the sample may be biased due to its dependence on the distribution obtained from the training set, it is possible that the sample value closest to the probability value of the new sample may also be biased. By assigning weight to the distance, we can obtain the deviation between this sample and other samples.

This section employs standard deviation to express the degree of deviation between new and existing samples. Markowitz's portfolio theory is one of the most widely used models among classic portfolio models. It measures risk using standard deviation and proposes the investment efficient frontier based on this measurement. Using the kernel regression method, this section calculates the standard deviation of the new sample based on the expected probability of fraud. The calculated fraud probability is then converted into the expected return. The kernel regression deviation of the sample is considered as the standard deviation for building the investment-efficient frontier of P2P, as shown in Figure 7.19. Investors can invest in the samples on the efficient frontier. There are approximately 160,000 samples that meet the efficient frontier, accounting for roughly half of the total number of samples.

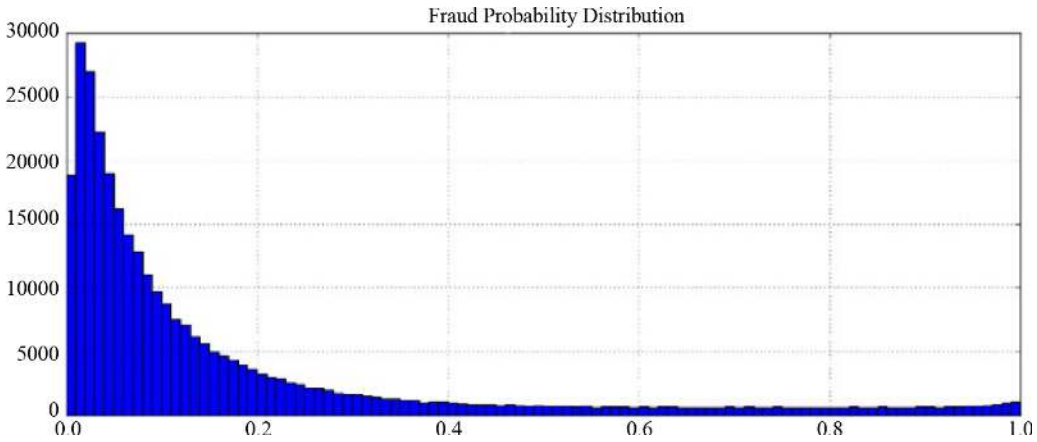


Figure 7.18. Fraud probability distribution.

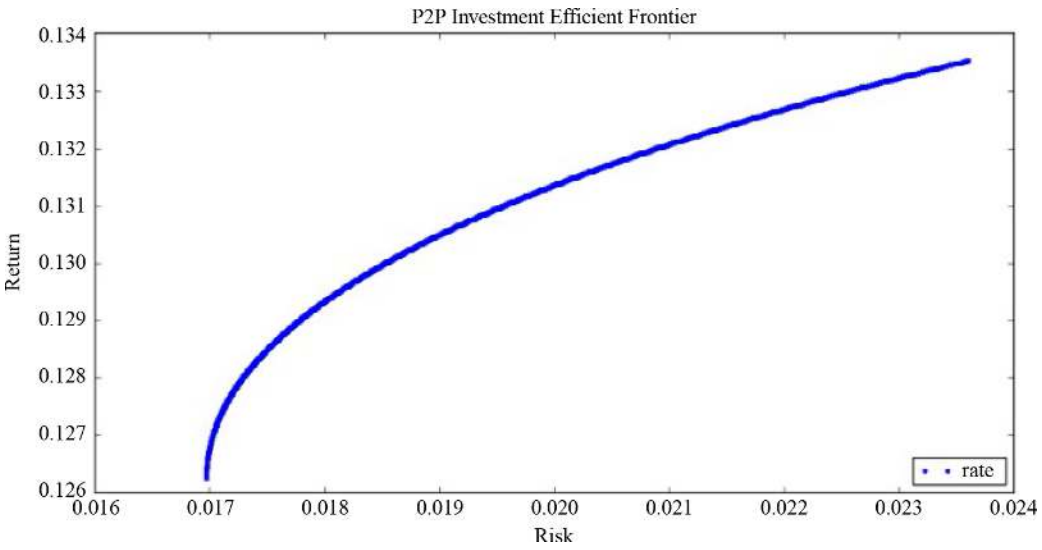


Figure 7.19. P2P investment efficient frontier.

7.6.2 *Algorithm simulation*

To assess the effectiveness of the kernel regression-based investment recommendation algorithm in P2P investment, this section conducts a simulation. We simulate the selection of investment targets in three ways: blind investment, investment based on a risk identification algorithm using random forest, and investment based on the recommendation algorithm using kernel regression. Each simulation involves making a total of 100 investments, with each investment consisting of investment in 100 loans. The resulting delinquency chart is presented in Figure 7.20.

Table 7.6 and Figure 7.20 demonstrate that nearly every blind investment includes delinquent loans, and the probability of fraud in blind investment varies greatly, with up to 9% of loans experiencing fraud. The random forest-based fraud detection algorithm has significantly improved blind investment. The highest fraud rate observed is 3%, and most simulation scenarios do not involve fraudulent loans. In comparison, the kernel regression recommendation algorithm has a lower proportion of fraudulent loans. The simulation shows that the maximum frequency of a fraudulent loan appearing is once. In most investment processes, no fraudulent loans occur. Finally, this section presents the results of simulations to calculate the fraud rates of blind investment, the random forest-based fraud detection algorithm, and the kernel regression

recommendation algorithm. The fraud rate of blind investment is 4%, while the fraud rate of the random forest-based fraud detection algorithm is 0.39%. The kernel regression recommendation algorithm, established based on the random forest fraud identification algorithm, has a significant improvement effect on investment, with a fraud probability of only 0.09%.

Table 7.6. Portfolio fraud.

	Blind Investment	Risk Identification Model	Investment Regression Model
Percentage of Fraud	4%	0.39%	0.09%

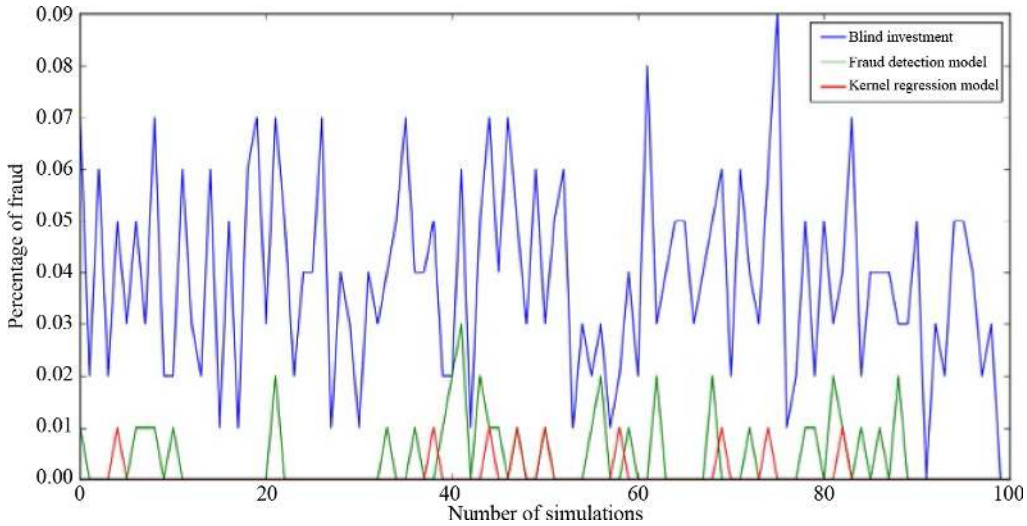


Figure 7.20. Simulation of fraud probability.

7.6.3 *Simulation considering reinvestment and algorithm performance and risk analysis under reinvestment*

Under the Paipaidai model, a loan is repaid monthly. If you do not reinvest the amount being repaid, this portion of funds will not receive any future returns. However, reinvesting will increase the risk of fraud in existing investments. Simulation shows that reinvestment amplifies the risk of fraud by 3–4 times. Therefore, when calculating returns, it is necessary to consider the expected returns if reinvestment occurs. We performed a reinvestment simulation on the previous investment results and obtained the reinvestment income chart shown in Figure 7.21.

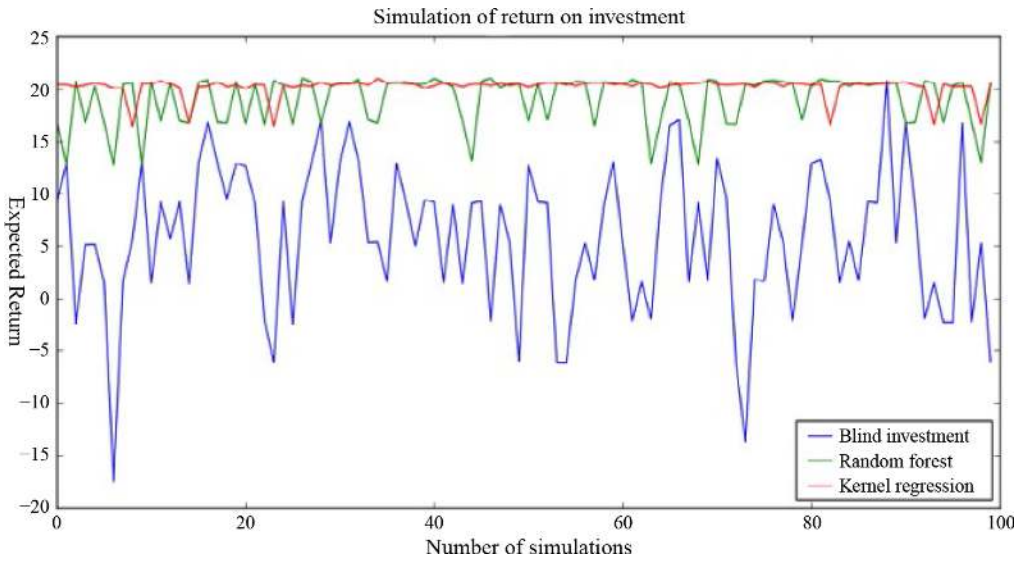


Figure 7.21. Simulation of return on investment.

Table 7.7. Expected return on investment portfolio.

	Blind Investment	Risk Identification Model	Investment Regression Model
Expected Return	5.8%	19.0%	20.2%

According to Figure 7.21, blind investing on the Paipaidai platform carries a high risk even if the platform controls the bad debt rate within 4%. Figure 7.22 shows the yield distribution of blind investment.

It can be seen that the investment return of blind investment is concentrated at about 10%, but the probability that other return scenarios will occur is not low, and even negative return scenarios may occur, and the expected return is 5.8% as shown in Table 7.7. After using the fraud detection algorithm, investment risks have been controlled. It can be seen that the investment return of the random forest fraud identification algorithm will fluctuate, while the fluctuation of the kernel regression method is smaller than that of the random forest algorithm. The distribution of returns for both is shown in Figure 7.23.

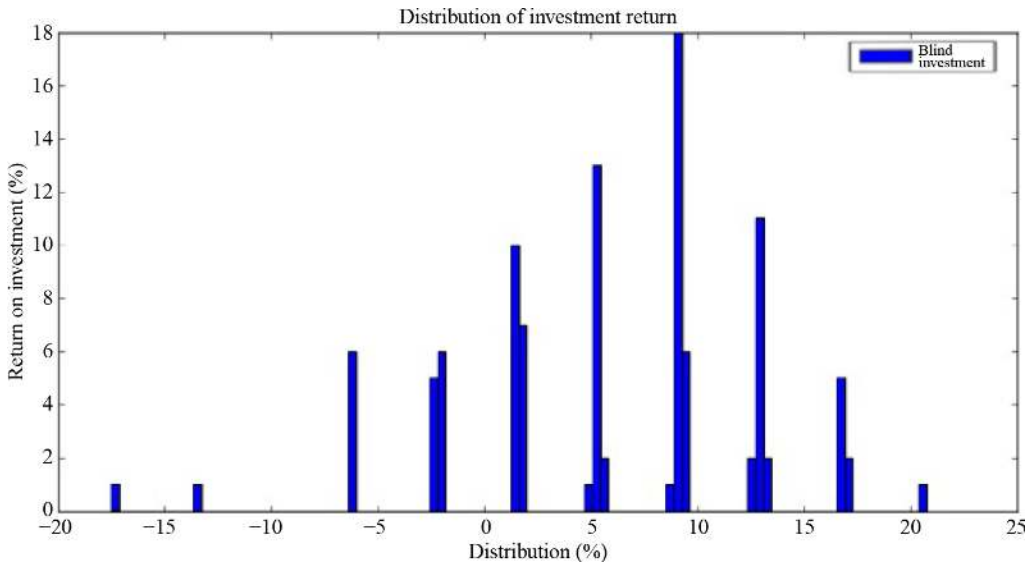


Figure 7.22. Blind investment return distribution.

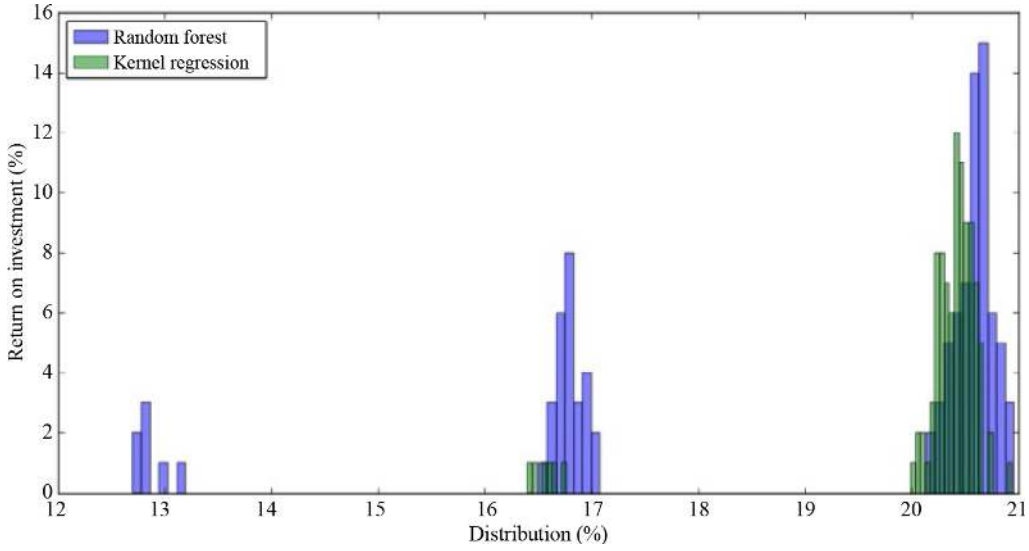


Figure 7.23. Comparison of investment return between random forest fraud detection algorithm and kernel regression recommendation algorithm.

It can be seen that the investment results of the kernel regression-based algorithm have more investment returns concentrated between 20% and 21% than the random forest-based fraud detection algorithm. However, the investment return of the fraud detection algorithm is, in some cases, greater

than that of the kernel regression method in the 20–21% return interval. The expected return on investment for the two algorithms is 19.0% and 20.2%, respectively.

7.7 Conclusion

In this chapter, we conduct research on P2P risk identification and corresponding investment models, using crawler technology and OpenAPI to obtain the latest real P2P data that reflect real China's investment environment. We used machine learning algorithms to study these data and found that variables such as educational background, borrowing history, and borrowing behavior have a significant impact on fraud probability. In terms of algorithm, we first proposed a fusion algorithm that combines the Apriori algorithm and the random forest algorithm and proved through empirical analysis that the performance of the fusion algorithm is significantly improved compared to the original algorithm. Secondly, based on the above fraud model, this chapter proposes an investment recommendation model that combines the kernel regression method with traditional financial theory and, based on this, calculates the effective investment frontier of P2P investment. The main work of this chapter is as follows:

(1) New methods are proposed to improve the current algorithm through machine learning methods, and the algorithm performance has been significantly improved.
(2) The underlying risk of each loan on P2P platforms is recharacterized through a kernel regression approach, and the efficient frontier of investment in P2P is calculated by integrating the Markowitz portfolio theory.
(3) Considering the possible risk amplification caused by reinvestment, the real investment environment is simulated as much as possible. The simulation results prove that the investment recommendation algorithm based on kernel regression can effectively improve investment returns.

Chapter 8

Risk Identification of Internet Financial Platform Based on Text Mining

8.1 Introduction

Internet finance is a new industry which has grown out of the rapid development of the Internet. Internet finance, which relies on the Internet for capital exchange and other activities, naturally has more information access channels than traditional finance. Taking online lending as an example, when choosing a platform, both lenders and borrowers can not only judge the quality of the platform through its basic information (such as size, years of establishment, etc.) but also refer to the evaluation of Internet users, and even further exchange information with them. Therefore, how to effectively use these evaluations has become a new topic in Internet finance research.

Comments, in the form of text, are a type of unstructured data. Compared to structured data, their structure is irregular, and there is no predetermined data pattern. Therefore, traditional data mining methods cannot obtain effective information from them and we will need to rely on advanced text minings methods to extract information from this type of unstructured data.

8.2 Text Mining Methods

Text mining, which extracts useful information from unstructured text, is a branch of data mining. However, this does not mean that simply applying data mining techniques to a large collection of text can achieve the goal of text mining. There is still a lot of preparation work to be done. The preparation work for text mining consists of three steps: text collection, text analysis, and feature pruning.

Text Collection: The text data that needs to be mined may have different types and are scattered in many places. Thus the first crucial step is to find and retrieve all those texts that are potentially relevant to the work at hand.

Text Analysis: Compared to structured data in databases, text has limited or no structure at all. In addition, the content of the document is a natural language used by humans, making it difficult for computers to process its semantics. The special characteristics of text data sources make existing data mining techniques unable to be directly applied to them. It is necessary to analyze the text and extract metadata representing its features, which can be saved in a structured form as an intermediate representation of the document. The aim of text analysis is to scan and extract the necessary facts from the text.

Feature Pruning: This includes both horizontal selection and vertical projection. Horizontal selection refers to removing noisy documents to improve mining accuracy or selecting only a portion of samples to improve mining efficiency when there are too many documents. Vertical projection refers to selecting useful features according to the mining target, and through feature pruning, an effective and concise subset of features representing the document set can be obtained. Based on this, various document mining works can be carried out.

After feature pruning, text mining can be carried out. Common text mining methods include text classification, text clustering, topic extraction, abstract extraction, and so on.

Text Classification: It refers to pre-classifying partial texts and training models and then classifying unclassified texts. Common text classification

scenarios include the following: (1) classifying the emotions of the text, such as positive emotions, neutral emotions, and negative emotions; (2) classifying the quality of text, such as spam and non-spam. Common classification methods include simple Bayesian classification, matrix transformation, *K*-nearest neighbor (KNN) reference classification algorithm, and support vector machine classification method.

Text Clustering: Clustering similar texts into one category based on their content is different from text classification in that it does not require prior judgment on certain parts of the text. The commonly used clustering methods include hierarchical clustering, plane partitioning, simple Bayesian clustering, KNN clustering, hierarchical clustering, and concept-based text clustering.

Topic Extraction: By calculating the importance of different words in a text, several words that best represent the text are extracted and combined to form the theme of the text. One of the most commonly used models is the LDA topic model.

Abstract Extraction: Similar to topic extraction, abstract extraction also extracts important vocabulary from the text. However, unlike topic extraction, the ultimate goal of abstract extraction is to combine important vocabulary into sentences that are concise, objective, comprehensive and have good readability, which are used as abstracts for readers to quickly browse. Based on the source of sentences used to generate abstracts, automatic summarization methods can be divided into two categories: one is to generate abstracts entirely using sentences from the original text, and the other is to automatically generate sentences to express the content of the document. Currently, most of the methods used involve extraction generation.

8.3 Text Mining in Internet Financial Platform

Texts involved in Internet financial platforms include textual descriptions of the platforms themselves, comments on the platforms by Internet users, and comments on the platforms by news commentators.

Fan Chao *et al*. (2017) earlier used text mining technology in the study of Internet financial platforms, which used data from 444 platforms to explore the characteristics of problematic platforms and crawled nearly 300,000 netizen comments, extracting the relevant features (such as online popularity and attitudes) through text mining technology.[96] They found that netizen comments play an important role in screening the risk of online lending platforms. However, Fan Chao *et al*. did not compare the effectiveness of the model before and after adding comments from netizens and only drew conclusions based on the importance of features related to comments from netizens in the model.[96]

8.4 Risk Identification of Internet Financial Platform

8.4.1 *Data sources and acquisition*

The data in this chapter mainly come from Wangdaizhijia, P2P eye, and CSMAR Solution. The data are obtained through crawling and downloading and are then sorted and stored in MySQL. The sources and acquisition methods of each type of data are shown in Table 8.1.

Among them, the basic information data of the platform includes basic information such as platform launching time, platform paid in capital, and operating information such as platform average revenue and platform management fees. The platform evaluation data of Wangdaizhijia includes evaluation content, evaluation scores, and the opinions of others on the evaluation content (useful or useless). The platform evaluation data of P2P eye includes evaluation content and classification (like, dislike, neutral). Industry news data include news titles and news content. The unemployment rate data is the quarterly data of the urban registered unemployment rate. The stock market index includes the Shanghai Composite Index, Shenzhen Composite Index, CSI 300 Index, and their

returns. The risk-free interest rate data is the one-year interest rate for overnight Certificates of Deposit.

Table 8.1. Data details.

Perspectives	Data	Source	Acquisition method	Data frequency	Time interval	Data volum
Platform level	Platform basic information	Wangdaizhijia	Crawling	Real time	2007–2019	6,504
	Platform evaluation	Wangdaizhijia	Crawling	Real time	2007–2019	207,96
	Platform evaluation	P2P eye	Crawling	Real time	2007–2019	314,88
Industry level	Industry news	Wangdaizhijia	Crawling	Real time	2007–2019	298,08
Macroeconomic level	Unemployment rate	CSMAR	Downloading	Quarterly	2007–2019	49
	Stock market index	CSMAR	Downloading	Daily	2007–2019	3,061
	Risk-free interest rate	CSMAR	Downloading	Daily	2007–2019	4,596

8.4.2 *Text mining*

8.4.2.1 *Sentiment analysis*

There are two main types of methods for sentiment analysis: sentiment dictionary-based and machine learning-based. As the vocabulary in the field of online lending is relatively new and there is no mature sentiment dictionary yet, it is difficult to capture the emotions that the commenters are trying to express by using the sentiment dictionary method, so the latter is used in this chapter.

Meanwhile, the texts involved in this chapter can be divided into two categories: (1) the comment texts of platform users, characterized by shorter text length, less standardized expression, and relatively simple emotions; (2) news texts, characterized by longer text length, relatively standardized expression, and relatively rich emotions. There are significant differences

between the two types of texts, so this chapter conducts a sentiment analysis on these two types of texts separately.

8.4.2.2 *Sentiment analysis of comment text*

Machine learning-based sentiment analysis is a supervised learning approach that requires training models on partially labeled data (such as negative and positive) and annotating unlabeled data with the trained model. For publicly available text corpora with high similarity (such as stock reviews and specific product reviews), models can be trained using publicly available corpora. However, for publicly available text corpora without high similarity, researchers need to manually annotate them using relevant domain knowledge.

The comment text data in this chapter falls between the two, i.e., there is no similar public corpus, but each user of P2P eye is required to rate when commenting. Therefore, it can be considered that the data crawled from P2P eye contains labels, as shown in Figure 8.1.

However, the quality of review data on P2P Eye, an online lending platform is far inferior to that of public corpus data, mainly due to the problem of some users commenting on multiple platforms with the same content. As shown in Figure 8.2, user “15101105179” commented, “Currently, investors are very enthusiastic about investing in wealth management products, and it is recommended to pay attention to product risks. Generally, product risks are evaluated from multiple dimensions, such as the background of the financing party, the transaction structure of the product, and whether there is a guarantee. I usually ask if the platform’s product is worth investing in through Haocaitou.” This comment appeared on up to 14 platforms, including Yirendai, Paipaidai, and Weidai.com. These types of comments are mostly meaningless content for judging the quality of the platform, and due to their repetition, they are not helpful for model training.



Figure 8.1. Online loan P2P Eye comment page.

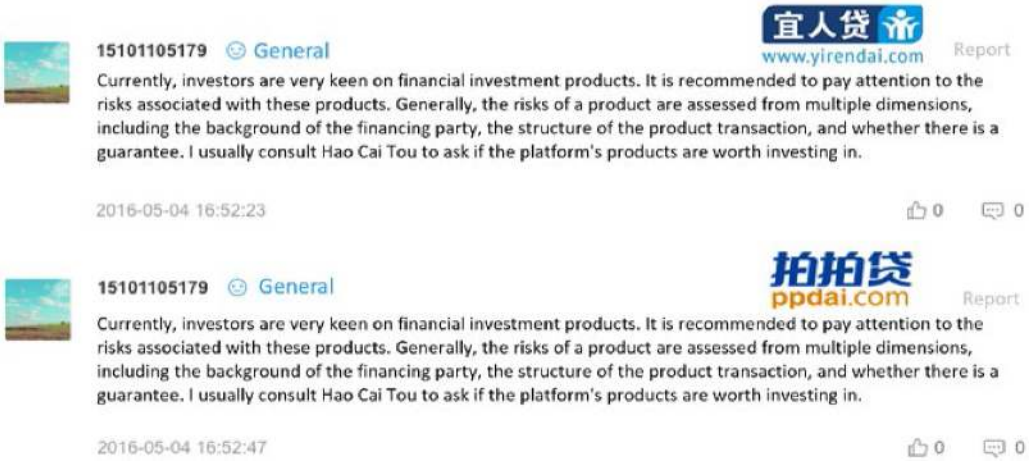


Figure 8.2. Repetitive comment text for the same user.

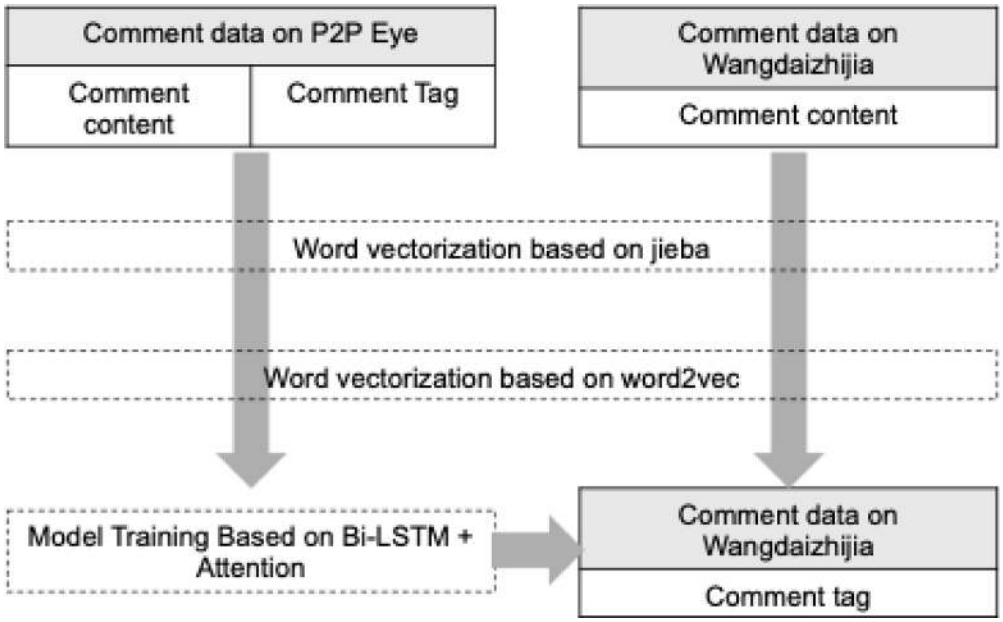


Figure 8.3. Sentiment analysis framework for comment texts.

Therefore, this chapter first deletes the duplicate comments and then uses the framework in Figure 8.3 to train the sentiment analysis model based on P2P eye and annotate the sentiment of Wangdaizhijia comment text.

8.4.2.3 *Sentiment analysis of news text*

Due to the existence of specialized terminology within each industry, such as "uninvested capital" in the online lending industry and "memory" in the computer industry, there are certain differences in relevant news texts between different industries. However, unlike comments from netizens, most news commentators use more standardized descriptive and evaluative sentences. So, despite the differences in the industries, the news texts of different industries are similar at the model level. Therefore, the sentiment analysis of news texts in this chapter uses Sogou News text corpora as training samples for model training and uses the trained model to classify the sentiment of news texts in this chapter, as shown in Figure 8.4.

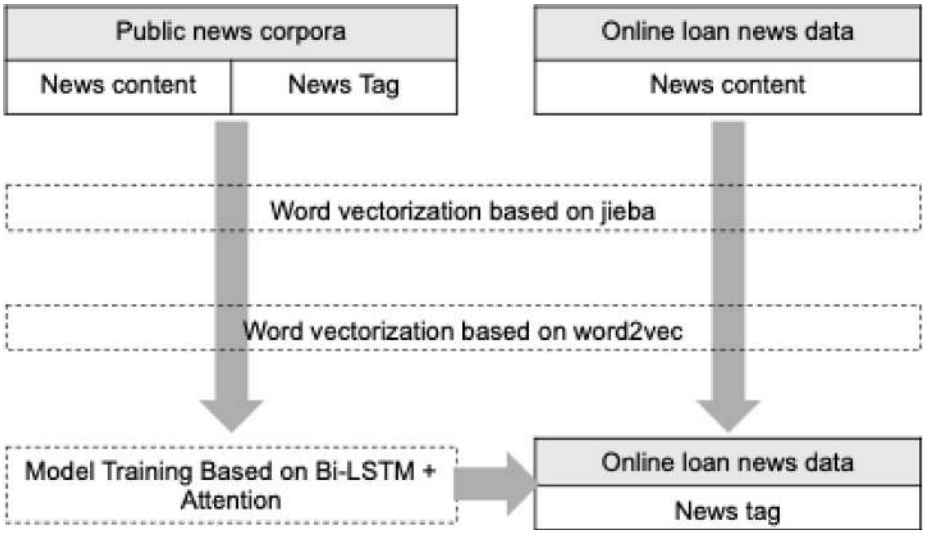


Figure 8.4. Framework for sentiment analysis of news texts.

8.4.2.4 *Topic extraction*

LDA is the most common topic extraction model, but since the comments in this chapter are all aimed at P2P platforms, the similarity between topics extracted by LDA alone is too high-level to distinguish well. Therefore, this chapter adopts a combination of TF-IDF and LDA methods to extract topics.

We have borrowed the idea of TF-IDF, which states that words that appear more frequently in a single document but appear less frequently in all documents are more important words, and vice versa. First, we remove the less important words from each comment, and then use LDA to extract themes from the processed comments. Finally, the following three themes are extracted:

Topic 0: Withdrawal+Receipt+Custody+Bank+Trust+Funds+Interest+Overdue+Good luck+Real

Topic 1: Good+Investment+Income+Payment Collection+Experience+Feeling+Hope+Friends+Getting Better+Stability

Topic 2: Platform+Activities+Financial Management+Running Away+Many+Red Envelopes+Projects+Observations+Popularity+Choices

By using the trained LDA model, the topic distribution of each comment can be calculated. If the comment “The platform is good and the secure payment experience is good!” has a topic distribution of {“Topic 0”: 0.0833371, “Topic 1”: 0.5833469, and “Topic 2”: 0.33331597}, among which Topic 1 has the highest probability, then the comment is labeled as Topic 1.

8.4.2.5 *Other texts*

In the platform overview of Wangdaizhijia, fields such as recharge fees and VIP fees are descriptive texts, such as “Quick payment fee rate: 0.15%, transfer recharge fees are subject to actual bank charges,” “No recharge fee,” “Temporary free for users, paid by the platform,” etc. This chapter converts keywords into 0/1 variables through keyword recognition. If there are keywords such as “free,” “temporarily unavailable,” or “not collected,” they are marked as 0. Otherwise, they are marked as 1 to determine whether each platform charges additional fees, such as recharge fees and VIP fees.

8.4.3 *Numerical data processing*

8.4.3.1 *Platform status*

We use the variable platStatus to capture the basic information of online lending platforms on Wangdaizhijia. platStatus is first divided into 10 categories, each assigned a numerical value the corresponding definitions of each value are shown in Table 8.2, where 1 (normal) indicates that the platform is still operating normally, and 7 indicates the platform is transitioning to a benign exit from the market. We then recoded the platStatus variable into a binary indicator. Platforms with status 1 are as labeled as normal (and marked as “0”), while platforms with statuses other than 1 or 7 are classified as fraudulent platforms and assigned a value of 1. Platforms with an original status of 7 are excluded from the sample, resulting in the removal of 79 platforms. After the above processing, 5,828 platforms with a platStatus of 1 and 626 platforms with a platStatus of 0 were obtained. However, due to insufficient data for platforms with a short operating history, which provides limited reference value, this chapter removed platforms with an operating time of less than 3 months (90 days)

and ultimately obtained 5,761 platforms with a platStatus of 1 and 612 platforms with a platStatus of 0.

Table 8.2. PlatStatus description.

PlatStatus	Description
1	Normal
2	Suspension of business
3	Withdrawal difficulties
4	Make off with money
5	Economic investigation intervention
6	Dispute
7	Transformation
8	Suspend bidding
9	Delayed redemption
10	Website shutdown

8.4.3.2 *Number of platforms*

Using information on Wangdaizhijia, we constructed two variables: onlineDate (platform launch date) and problemTime (date when the platform started to having issues). For the purpose of subsequent analysis, a new variable, "daily active platform quantity," is defined, which refers to the number of platforms that are between onlineDate and problemTime on such day (normal operating platforms). According to Wangdaizhijia, the earliest recorded onlineDate for online lending platforms is June 17, 2007. After traversing from June 17, 2007, to September 1, 2019, the number of daily active platforms can be obtained for subsequent analysis.

8.5 Fraud Identification on Online Lending Platforms

On the basis of reviewing the literature, this chapter proposes including comment data, industry data, and macro data in the research to improve the fraud identification rate of China's online lending platforms. To explore the effectiveness of the variables proposed in this chapter, we will group the

data and gradually add the three types of variables proposed in this chapter based on the commonly used variables in the literature. We will compare and analyze whether the added variables are helpful in improving the fraud identification rate of China's online lending platforms. At the same time, by exploring the direction and magnitude of the effects of various variables on the possibility of fraud on China's online lending platforms, key factors that lead to fraud on online lending platforms are identified, providing a certain basis for regulatory oversight.

8.5.1 *Feature construction*

Two types of data can be obtained through data preprocessing: (1) static data, which either does not change over time or can only be obtained in its final state, such as the platform's paid-in capital and the city where the platform is located; (2) dynamic data, which refers to data that changes over time, such as comments, news, and unemployment rates.

However, existing models are unable to handle mixed static and dynamic data separately. So, it is necessary to expand static data into dynamic form or compress dynamic data into static form. Due to the fact that this chapter is a static prediction for the final state of the platform, the dynamic data are compressed into static form based on the operating period of each platform.

8.5.1.1 *Mean features*

For quarterly urban unemployment rates, the operating period of a single platform may cover a non-integer number of quarters, so a weighted average method is adopted for calculation. As shown in Figure 8.5, the launch date of a certain platform is August 20, 2015, and the problem time is April 30, 2017. Therefore, the weight of Q3 in 2015 is 4/9, and Q2 in 2017 is 1/3. The weight of other quarters is 1. Based on this, the weighted average urban unemployment rate for each quarter is obtained to obtain the average urban unemployment rate unemployment_rate corresponding to the platform.

For daily risk-free interest rates, stock market returns (using the CSI 300 Index as an example), number of platforms, and number of comments, the operating period of a single platform covers an integer number of

natural days. Therefore, a simple average method is used to calculate the average risk-free interest rate risk_free_rate, average stock market return stock_index_rate, average number of platforms plat_num, and daily average number of comments corresponding to the platform.

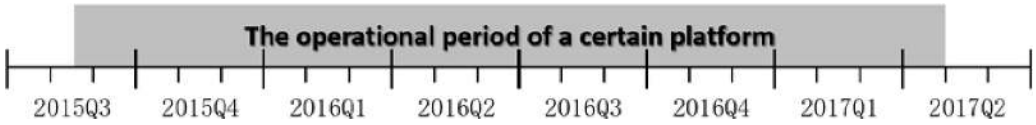


Figure 8.5. Example of calculating the urban unemployment rate.

8.5.1.2 *Proportion characteristics*

For data such as comment themes, comment emotions, and news emotions, the amount of comment data corresponding to the operation period of a single platform varies, and it is not possible to obtain the distribution of topics and emotions for each platform by comparing the absolute data volume of each platform. Therefore, the method of calculating the proportion is used to obtain relevant characteristics. Taking emotion_1 as an example, it refers to "the number of positive comments on a certain platform/the number of all comments on that platform," and similarly we constructed theme_0, theme_1, theme_2, emotion_1, emotion_0, emotion_-1. Taking news_1 as an example, it refers to "the number of positive industry news during the operation period of a certain platform/the number of all industry news during the operation period of the platform." Similarly we constructed, news_0, and news_-1.

As a result, the 28 variables in Table 8.3 were ultimately obtained.

To compare the effectiveness of the variables proposed in this chapter, the 28 variables from the above four perspectives were divided into the following eight groups based on their perspectives:

Feature_1: Platform level — Structural.

Feature_2: Platform level — Structural + Platform level — Non-Structural.

Feature_3: Platform level — Structural + Industry level.

Feature_4: Platform level — Structural + Macro level.

Feature_5: Platform level — Structural + Platform level — Non-Structural + Industry level.

Feature_6: Platform level — Structural + Platform level — Unstructured + Macro level.

Feature_7: Platform level — Structural + Industry level + Macro level.

Feature_8: Platform level — Structural + Platform level — Non-Structural + Industry level + Macro level.

Table 8.3. Fraud identification variables and description.

Variable perspective		Variable name	Meaning of variables	Variable type
Platform level	Structural type	ActualCapital	Paid-up capital	Numerical value
		OnlineTimeStr-num	Years of online operation	Numerical value
		PlatEarningsCode	Platform revenue	Numerical value
		AutoBidCode	Support automatic bidding or not	0–1
		Association	Join the regulatory association or not	0–1
		BindingFlag	Is there a bidding guarantee?	0–1
		BankFunds	Whether the funds are under custody	0–1
		Inspection	Is there a guarantee agency available?	0–1
		ManageExpense-01	Is a management fee charged?	0–1
		RechargeExpense-01	Is a recharge fee charged?	0–1
		VipExpense-01	Is a VIP fee charged?	0–1
		LocationCity	City code	Category
		NewTrustCreditorCode	Debt transfer	Category
		PlatBG	Platform background	Category
		Theme0	Proportion of comments with	Numerical

			a theme of 0 during operation	value
		Theme1	Proportion of comments with a theme of 1 during operation	Numerical value
		Theme2	Proportion of comments with a theme of 2 during operation	Numerical value
	Unstructured type	Emotion_1	Proportion of positive reviews during operation	Numerical value
		Emotion_0	Proportion of neutral reviews and comments during operation	Numerical value
		Emotion_-1	Proportion of negative reviews during operation	Numerical value
		Comment_num	Daily average number of comments during operation	Numerical value
		platform_num	Average number of platforms during operation	Numerical value
		News_1	Proportion of positive news during operation	Numerical value
Industry level				
		News_0	Proportion of neutral news during operation	Numerical value
		News_-1	Proportion of negative news during operation	Numerical value
		Risk_free_rate	Average risk-free interest rate during operation	Numerical value
Macro level		Unemployment_rate	Average urban unemployment rate during operation	Numerical value
		stock_index_rate	Average stock market return during operation	Numerical value

Group 1 is the base group, which only contains basic information about each platform and serves as the control group to confirm the effects of other variable groups. Groups 2–4 add comments-related variables from each platform, industry-level-related variables, and macro-level-related variables

on the basis of the first group. Groups 5–7 add any two of the other three types of variables on the basis of Group 1. Group 8 includes all the variables involved in this chapter.

8.5.2 *Fraud identification*

8.5.2.1 *Model training*

This chapter selects seven commonly used classification models for fraud identification on online lending platforms for training, including Naive Bayes (NB), *K*-nearest neighbors (KNN), logistic regression (LR), support vector classification machine (SVC), decision tree (DT), random forest (RF), and Gradient Boosting Decision Trees (GBDT), and also five commonly used indicators of classification models from Table 8.4 as the classification evaluation criteria for each model and each group of variables.

At the same time, in order to confirm the robustness of each model, i.e., the predictive ability outside the sample rather than within the sample, the data are divided in a ratio of 7:3 between the training set and the test set in the model. The training set is used for training, and the model performance is evaluated using the test set. After partitioning, the ratio of positive and negative samples in the training set is 4015:421, and the ratio of the positive and negative samples in the test set is 1746:191.

The results of each indicator on the test set of models trained on different data are shown in Figures 8.6–8.10.

8.5.2.2 *Comparison of performance indicators*

This chapter selects five performance indicators, among which AUC and accuracy are overall performance indicators, accuracy and recall are partial performance indicators, and F1 Score considers both overall and partial performance.

For AUC and accuracy, both evaluate the model as a whole, but the evaluation criteria are different. Accuracy is determined by the optimal threshold, wherein each sample is assigned to the category with the highest likelihood. Conversely, AUC is computed across all thresholds, that is, for each sample, it considers the probabilities across all possible categories. Therefore, compared to accuracy, AUC is more suitable for non-equilibrium

samples.[97] The accuracy of each model and data combination in this chapter is around 0.9. However, due to the class imbalance problem of the data in this chapter (more positive samples than negative samples), according to the calculation formula of accuracy, even if the model predicts all samples as positive samples, its accuracy can still reach 0.9 which is not helpful for identifying fraudulent platforms. Therefore, this chapter will not evaluate the model and data based on accuracy. Although AUC is more suitable for non-equilibrium samples compared to accuracy, it is still affected by data imbalance. Therefore, the following paragraphs will continue to explore the applicability of other evaluation indicators to find the most suitable evaluation indicator.

Table 8.4. Model evaluation indicators and meanings.

Index		Indicator calculation	Meaning of indicators
Based on overall	Accuracy	(TN+TP)/(TN+TP+FN+FP)	Proportion of correctly identified samples in the aggregate samples.
	AUC	Area under ROC curve	Performance evaluation of models based on true positive rate and false positive rate.
Based on partial	Precision	TP/(TP+FP)	Proportion of samples classified as positive samples that are actually positive samples.
	Recall rate	TP/(TP+FN)	Proportion of the actual positive instances that are correctly identified as positive in the sample set.
Base on overall and partial	F1 Score	2 * Precision * Recall/(Precision+Recall)	Harmonic mean of precision and recall.

Conceptual exposition
Positive example: A sample with platStatus 1, i.e., a platform where fraud issues have occurred.
Negative example: A sample with a platStatus of 0, i.e., a platform that has not experienced any fraud issues.
TN: True Negative, i.e., the model predicts the sample to be negative which in reality is indeed negative.
FP: False Positive, i.e., the model predicts the sample to be positive however in reality it is negative.
FN: False Negative, i.e., the model predicts the sample to be negative however in reality it is positive.
TP: True Positive, i.e., the model predicts the sample to be positive which in reality is indeed positive.
ROC curve: A curve plotted with false positive rate (FPR=FP/(FP+TN)) as the *x*-axis and true rate (TPR=TP/(TP+FN)) as the *y*-axis.

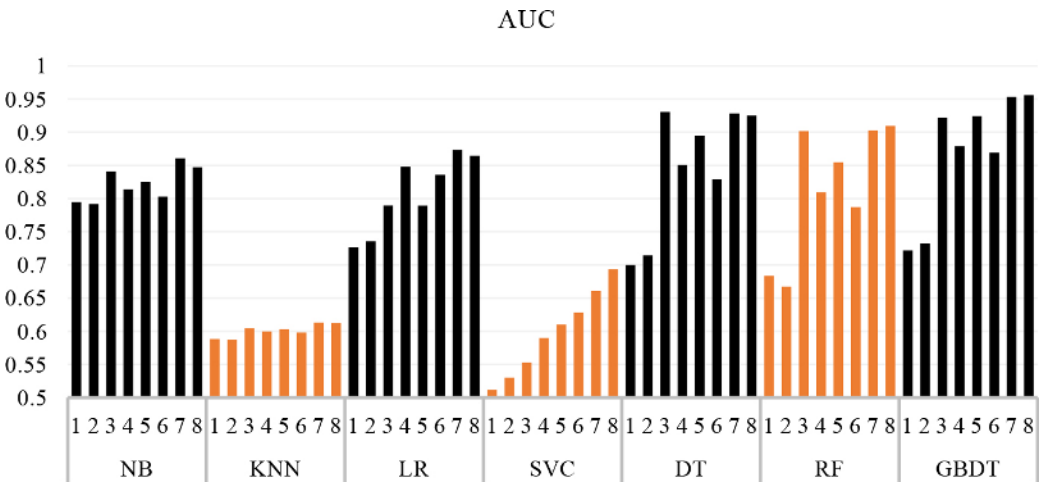


Figure 8.6. AUC of various data and model combinations trained on the test set.

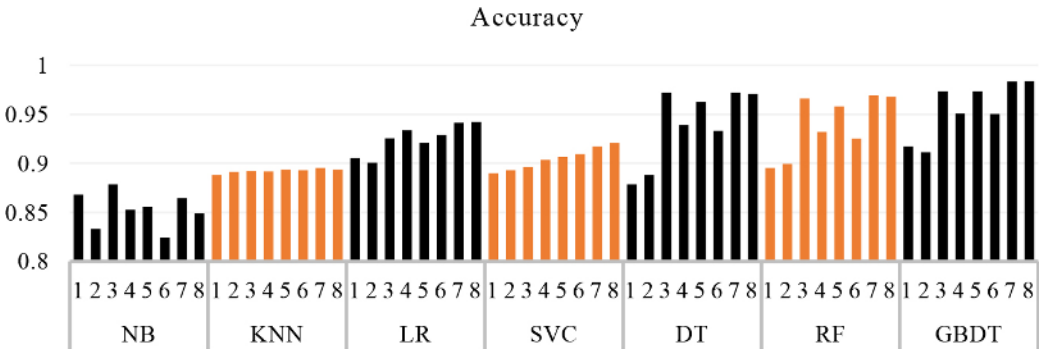


Figure 8.7. Accuracy of each data and model combination trained on the test set.

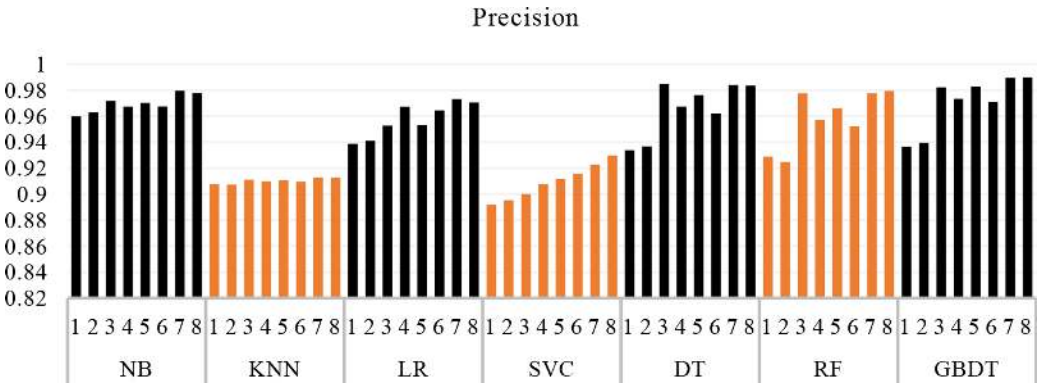


Figure 8.8. Precision of each data and model combination trained on the test set.

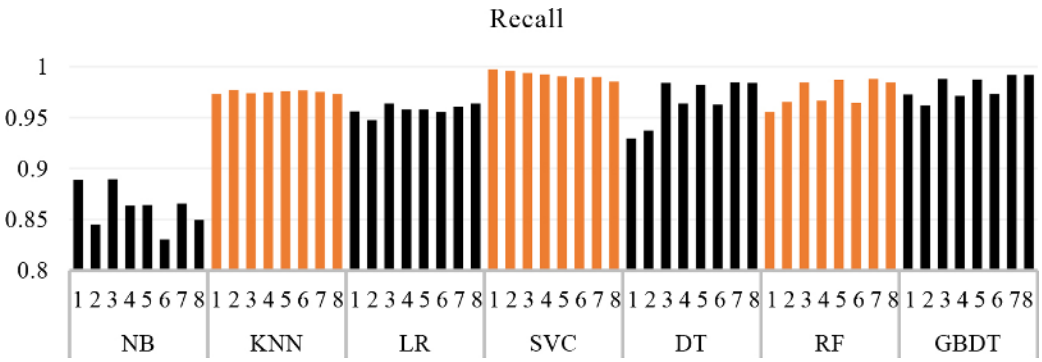


Figure 8.9. Recall on the test set after training various data and model combinations.

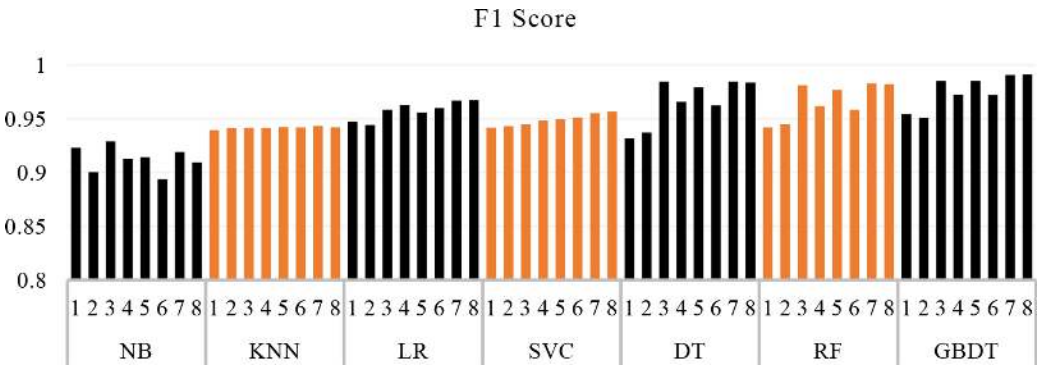


Figure 8.10. F1 Score on test set after training with various data and model combinations.

Accuracy and recall are both based on a partial evaluation of the model, but they focus on different aspects. Precision measures the proportion of true positive instances within the identified samples, equating to the fraction of fraudulent platforms correctly classified by the model. In contrast, recall measures the proportion of actual positive instances that are correctly classified, representing the model's ability to detect all fraudulent platforms. For balanced samples, the most suitable indicator can be selected between the two according to actual needs. If the objective is to ensure that all identified fraudulent platforms are indeed fraudulent, even at the expense of missing some actual fraudulent cases, accuracy is an appropriate evaluation metric. Conversely, if the priority is to identify all fraudulent platforms without concern for misclassifying non-fraudulent ones as fraudulent, the recall rate is the preferred metric. Analysis of Figures 8.8 and 8.9 reveals that various data and model combinations exhibit strong performance on

both indicators. However, for this chapter, the objective is to identify as many fraudulent platforms as possible without misjudging non-fraudulent platforms as fraudulent platforms. Therefore, it is necessary to find more suitable evaluation indicators.

F1 Score evaluates the model based on both partial and overall performance. This indicator is the harmonic mean of accuracy and recall, which can simultaneously consider both accuracy and recall, ensuring that the model can recognize as many positive samples as possible and "misunderstand" as few negative samples as possible. This characteristic is consistent with the objective of this chapter, so the F1 Score will be the main evaluation indicator in the following section.

8.5.2.3 *Model comparison*

From the perspective of AUC: DT and integrated models based on DTs, such as RF and GBDT, perform the best with an average of approximately 0.8; NB and LR come second, with a mean of approximately 0.7; KNN and support vector machines perform the worst, with an average of about 0.6.

From an accuracy perspective: LR and tree models perform the best, with mean values exceeding 0.9; KNN and support vector machines come in second place, with an average of approximately 0.9; NB performs the worst, with an average of approximately 0.83.

From the perspective of precision: NB, LR, and tree models perform the best, with mean values above 0.9; KNN and support vector machines perform poorly, with an average of about 0.9.

From the perspective of recall rate: All models except NB perform well, with an average of 0.95 or above; NB performs poorly, with an average of approximately 0.85.

From the perspective of F1 score: LR and tree models performed the best, with mean values exceeding 0.95; KNN and support vector machines come in second place, with an average slightly lower than 0.95; NB performs the worst, with an average of about 0.9.

In summary, the tree models have the best performance in identifying fraud on online lending platforms. RF and GBDT as integrated models have

better performance than single DTs. Compared to other methods, the integration effect of GBDT based on Boosting is slightly better than that of RF based on Bagging. Therefore, the following section will mainly use the GBDT method to explore variable comparison.

8.5.2.4 *Data comparison*

To explore the effectiveness of adding data from different levels (platform level — unstructured, industry level, macro level) in this chapter, the F1 Score of GBDT on the training set is shown in Figure 8.11. From the graph, it can be seen that based on the first set of data, which only uses platform-level structural variables for identifying fraudulent platforms, except for the second set of data, the F1 Score performance of all other data on the training set has improved. The following paragraphs will discuss the effectiveness of different data combinations in identifying fraudulent platforms.

For the second set of data, which includes platform-level unstructured variables on top of platform-level structured variables, its performance is significantly inferior to the first set of data. This indicates that the characteristics related to investor comments, including the emotions, themes, and quantity of comments, are not helpful for fraud identification on China's online lending platforms and even reduce the fraud identification rate. This discovery can, to some extent, reflect the unsophisticatedness of individual investors in China's online lending industry. Individual investors' comments on the platform cannot help other investors judge its quality.

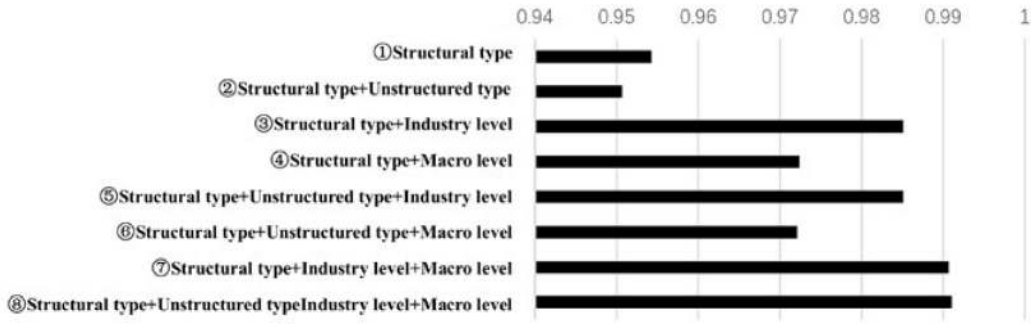


Figure 8.11. F1 Score on training set after GBDT training.

For the third group of data, which includes industry-level variables on the basis of platform-level structural variables, its performance is significantly better than the first group of data. At the industry level, there are two types of variables: one is related to industry news, such as the sentiment of industry news, and the other is the number of platforms in the industry. The addition of these two types of variables significantly improves the model's recognition rate for fraudulent platforms from 0.954 to 0.985. This indicates that in addition to the conditions of each platform itself, the overall performance of the industry is also crucial for the survival of individual platforms.

For the fourth group of data, which includes macro-level variables on the basis of platform-level structural variables, its performance is significantly better than the first group of data. Macro-level variables include risk-free interest rates, urban unemployment rates, and stock market returns. The addition of these three variables increases the recognition rate of the model for fraudulent platforms from 0.954 to 0.972. Although the increase is not as significant as the third group of variables, the results are also relatively objective, indicating that in addition to the conditions of each platform itself, the macroeconomic condition is also important for the survival of individual platforms, but the importance is slightly lower than the overall performance of the industry.

For the fifth group of data, which includes industry-level variables on the basis of the second group of data, or in other words, adding comments-related variables on the basis of the third group of data, its performance is significantly better than the first and second groups of data, but it is basically consistent with the performance of the third group, even slightly lower than the third group (from 0.98507 to 0.98506). This indicates that combining platform-level non-structural variables with industry-level variables can effectively improve the identification rate of fraudulent platforms, but the former has not made a contribution. This discovery further confirms the conclusion that the comments were not helpful in improving the identification rate.

For the sixth group of data, which includes macro-level variables on the basis of the second group of data, or in other words, adding relevant variables for comments on the basis of the fourth group of data, its performance is significantly better than that of the first and second groups of data, but it is basically consistent with the performance of the fourth

group, and even slightly lower than that of the fourth group (from 0.97236 to 0.97209). This indicates that combining non-structural variables at the platform level with macro-level variables can effectively improve the identification rate of fraudulent platforms, but the former has not made a contribution. This discovery further confirms the conclusion that comments were not helpful in improving the identification rate.

For the seventh group of data, which includes industry and macro level variables on the basis of platform level structural variables, its performance is significantly better than the first group of data, as well as the third and fourth groups. This indicates that the combination of industry-level and macro-level variables can effectively improve the identification rate of fraudulent platforms, and the combination effect of both is better than the effect of adding industry or macro level alone. That is, considering the overall level of industry and macroeconomy comprehensively can better identify fraudulent platforms.

For the eighth group of data, which includes industry-level unstructured variables on top of the seventh group, its performance is slightly better than that of the seventh group. This indicates that although adding relevant variables for comments alone does not help improve the recognition rate of online lending platforms, adding them on the basis of existing industry and macro data has a certain improvement effect. However, this increase is relatively small (from 0.9907 to 0.9910), and compared to the workload of adding comments (including data crawling, comment sentiment analysis, comment topic extraction, etc.), the cost-effectiveness is not high.

In summary, the effectiveness ranking of the data from different levels (platform level — unstructured, industry level, macro level) included in this chapter is industry level > macro level > platform level — unstructured. The unstructured data at the platform level can only slightly improve the recognition rate of fradulent online lending platforms when industry and macro-level data are added simultaneously, so it can be considered as ineffective data.

8.5.2.5 *Analysis of variable importance*

According to the previous discussion, it can be seen that the variables added in this chapter are helpful in improving the fraud identification rate of China’s online lending platforms. In order to further explore the role of

individual variables, we examine each variable from the perspective of both its importance and the direction of its influence within the model.

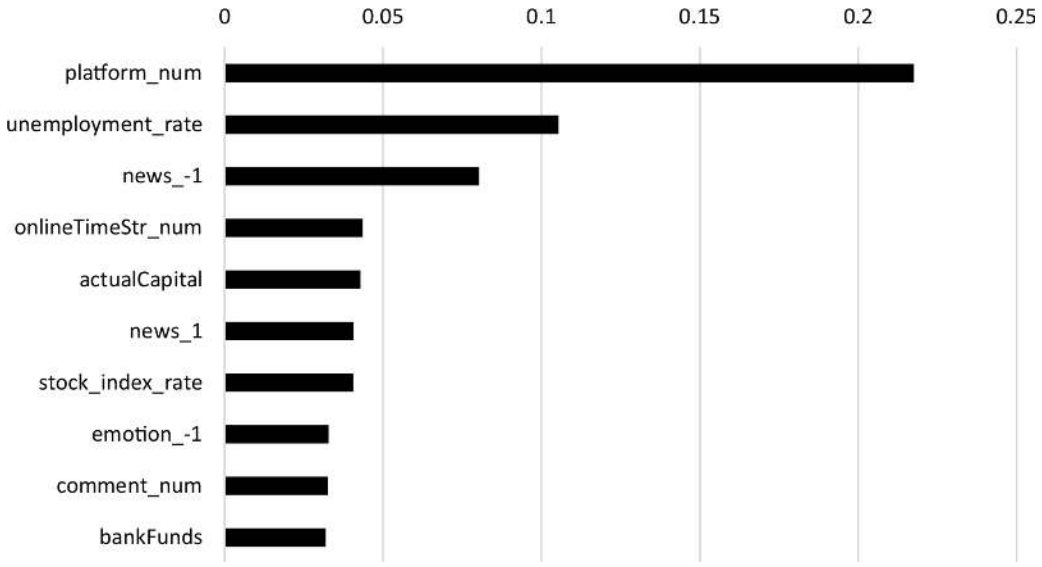


Figure 8.12. Variable importance based on GBDT.

For a well-trained tree model, the importance of variables can be obtained by calculating the reduction in weighted impurity of all non-leaf nodes during splitting. The more reduction, the more important the variable is. Therefore, utilizing this characteristic of the tree model, this chapter calculates and selects the top 10 variables ranked by importance based on the GBDT trained in the previous section, as shown in Figure 8.12.

From the graph, it can be seen that the most important variable is the number of platforms during the same period (platform_num), which is about twice as important as the second important variable — the average urban unemployment rate (unemployment_rate). Therefore, it can be seen that competition between platforms is a key factor affecting platform survival.

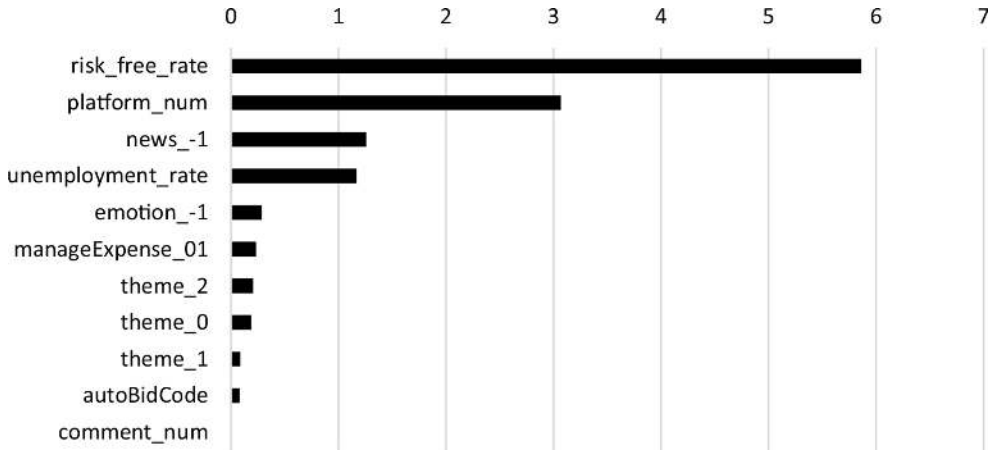
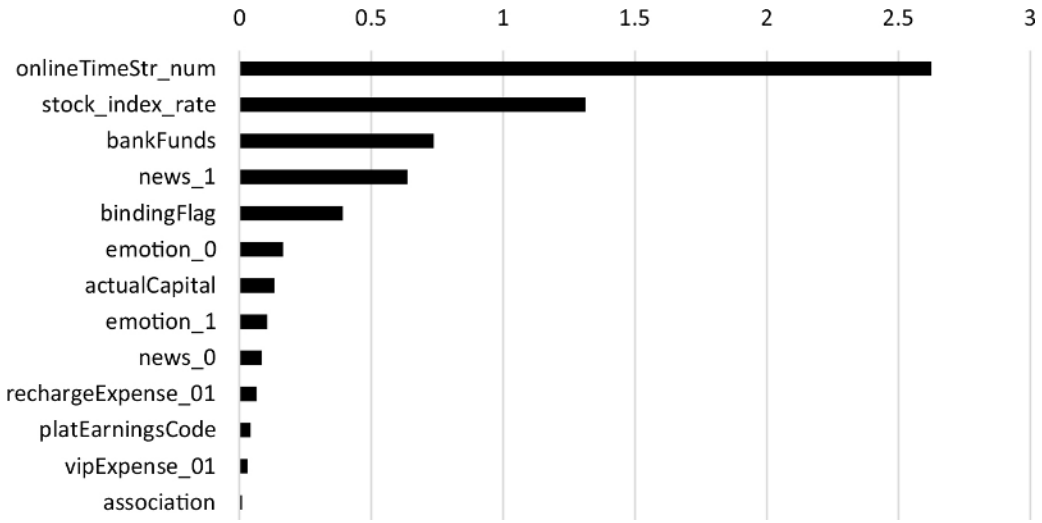


Figure 8.13. Variables with positive coefficients and their coefficients.



Figures 8.14. Variables with negative coefficients and their absolute values of coefficients.

Among the top 10, variables that also belong to the industry level similar to the number of platforms during the same period include the proportion of negative news (news_-1) and the proportion of positive news (news_1), while the proportion of neutral news (news_0) did not make the list. This indicates that compared to industry news without clear emotional tendencies, industry news with clear emotional tendencies has a greater role in determining whether platforms will have fraud problems.

The average urban unemployment rate and average stock return rate ranked second and seventh, respectively, while the macro-level variable average risk-free rate did not make the list. This indicates that the urban unemployment rate is more helpful for predicting fraud platforms, followed by the stock return rate, while the risk-free rate is the weakest.

As discussed in the previous section, platform-level unstructured variables are not effective variables for identifying fraudulent platforms. However, from the graph, it can be seen that the proportion of negative comments (emotion_-1) and the number of comments (comment_num) related to unstructured data rank eighth and ninth, respectively. Although the ranking is relatively low, it can still reflect that the negative comments and the number of comments are helpful in predicting platform fraud.

At the platform level, among the structural variables, the number of years the platform has been online (onlineTimeStr_num), actual paid-in capital and whether funds are under custody (bank funds) are also among the top 10, indicating that these three factors are also of great help in identifying fraudulent platforms.

8.5.2.6 *Analysis of impact direction*

Although variable importance analysis can identify variables that play a key role in identifying fraud on online lending platforms, it cannot determine the direction of each variable's impact, i.e., it cannot determine whether each variable has a promoting or inhibitory effect on the platform's fraudulent behavior. To address the above issues, Figures 8.13 and 8.14 show the coefficients of each variable corresponding to the model trained on the eighth set of data and LR.

According to the definition of LR, the positive or negative coefficients of the variables correspond to the direction of influence on the likelihood of platform fraud, with positive indicating promotion and negative indicating inhibition. The significance of the absolute value of the coefficient corresponding to a variable is the magnitude of its impact. For variables with positive coefficients, the larger the absolute value of their coefficients, the more likely the platform is to engage in fraudulent behavior. For variables with negative coefficients, the larger the absolute value of their coefficients, the less likely the platform is to engage in fraudulent behavior.

Figure 8.13 shows variables that have a promoting effect on the likelihood of fraud on platforms, including average risk-free rate (risk-free_rate), number of platforms during the same period (platform_num), proportion of negative news (news_-1), average urban unemployment rate (unemployment_rate), and proportion of negative reviews (emotion_-1).

First, the average risk-free interest rate has the strongest promoting effect, meaning that the higher the average risk-free interest rate during platform operation, the higher the likelihood of fraud issues occurring on the platform. The risk-free interest rate selected in this chapter is the one-year interest rate for overnight CDs. When other conditions remain unchanged (i.e., the average return of the platform remains unchanged), an increase in risk-free interest rates indicates an increase in opportunity costs for investing on the platform and a decrease in returns for investing on the platform. If the risk premium of investors in the online lending market remains unchanged, their expected return rate will exceed the average return of the platform, and investors are likely to choose not to invest temporarily. The increase in risk-free interest rates has the opposite effect on borrowers compared to investors, so borrowers are more likely to choose to shift from traditional lending markets to online lending markets. For the online lending market, the operational model of the platform determines that the platform must continuously invest funds to ensure continuous lending. The rise in risk-free interest rates disrupts this balance, which can exacerbate fraudulent behavior such as platforms choosing to default.

Secondly, the effect of the number of platforms during the same period is also very significant, i.e., the more competitors there are during the operation of the platform, the higher the likelihood of fraud problems occurring on the platform. The increase in competitors will lead to oversupply in the online lending market, where the number of online lending platforms far exceeds the demand of investors or borrowers with funds to borrow. On the one hand, in order to attract investors and borrowers in the online lending market, the platform needs to continuously launch products that are high-yield and high rebate products, resulting in the platform having losses in the early stage. Once the loss cannot be compensated within a certain period of time, the platform can only choose to exit the market. On the other hand, the situation of “supply exceeding demand” will inevitably lead to some platforms not being able to obtain enough users to maintain their daily expenses, and it will also lead to the

platform's exit. At the same time, due to the long benign exit cycle and high cost of online lending platforms (requiring the repayment of all bids on the platform or the platform to repay the full amount of investors who have not yet recovered their investment) and the imperfect punishment measures in the online lending market, most platforms that choose to exit choose to exit maliciously, including shutting down without paying investors back.

Thirdly, the proportion of negative news has a similar promoting effect as the average urban unemployment rate, i.e., the more negative news there is in the industry during platform operation and the higher the average urban unemployment rate, the higher the likelihood of fraud issues occurring on the platform. For the former, this result indicates that industry news has a certain predictive effect on the development of the industry. From an economic perspective, an increase in the average urban unemployment rate means that more labor resources cannot be effectively utilized, leading to an increase in unemployment and a decrease in total social demand. The driving force of economic growth will also weaken (Zeng, 2018)[148]. Therefore, the increase in the urban unemployment rate reflects a decrease in the economic growth rate. This result indicates the homogeneity between the online lending industry and macroeconomics, i.e., the decline of macroeconomics will lead to the decline of the online lending industry.

Finally, variables related to comments (including the proportion of negative comments, the proportion of different topics in comments, and the number of comments), whether management fees are charged (manageExpense-01) and whether automatic bidding (autoBidCode) is supported also have a certain promoting effect on the possibility of platform fraud, but the effect is not significant and will not be further explored in this chapter.

Figure 8.14 show variables that have a suppressive effect on the likelihood of fraud on the platform, including the number of years the platform has been online (onlineTimeStr num), average stock return rate (stock_index_rate), whether funds are under custody (bankFunds), and the proportion of positive news (news_1).

First, the number of years a platform has been online has the strongest inhibitory effect, meaning that the longer the platform is online, the lower the likelihood of fraud occurring on the platform. On the one hand, the longer the history of a platform, the richer its platform operation

experience, the more reasonable its fund utilization and management, and therefore, the less likely it is to encounter fraud problems. On the other hand, in the high-risk online lending industry, investors not only pursue returns but also pay great attention to risk control, including choosing well established platforms. Therefore, the longer the history of the platform, the more likely it is to be favored by investors and borrowers due to "established brand," resulting in more capital flow and being less prone to fraud. Second, the inhibitory effect of the average stock return rate is also more significant, i.e., the higher the average stock return rate, the lower the likelihood of platform fraud. This conclusion is inconsistent with the conclusion in literature possibly due to the selection of the average return rate of the CSI 300 Index as the proxy variable for the average stock return rate in this chapter, while the paper does not specify the specific index.[98] It is possible that other indices or a combination of multiple indices were selected, leading to inconsistent conclusions.

Again, the inhibitory effects of whether funds are entrusted and the proportion of positive news are similar. The custody of funds is a 0/1 variable, where a bankFunds value of 1 indicates that the platform's funds have a custody mechanism, and a value of 0 indicates that the platform's funds do not have a custody mechanism. Fund custody refers to the operation of fund flow in a third-party custody company without passing through the platform's bank account, which allows investors and borrowers to open accounts in the third-party payment platform's system. When a transaction occurs, funds are directly transferred from the investor's account to the borrower's account without passing through the platform layer.[a] Fund custody prevents the platform from misappropriating investor funds even in the event of poor operations, thereby suppressing the possibility of fraud on the platform. Meanwhile, the higher the proportion of positive news, the lower the likelihood of platform fraud, indicating that industry news has a certain predictive effect on the development of the industry.

Finally, other variables, including platform level and industry level, also have a certain inhibitory effect, but the effect is not significant and will not be further explored in this chapter.

8.5.2.7 *Summary*

Based on the above analysis, the following conclusions can be drawn in this section:

(1) **The tree model is more suitable for fraud identification on China's online lending platforms:** Compared to non-tree models such as LR, the tree model has better fraud identification performance, indicating the existence of a non-linear relationship between the indicators selected in this chapter and the platform status. The tree model can better capture this relationship, thus achieving better results.
(2) **The occurrence of fraud on platforms is closely related to the online lending industry and macroeconomic factors:** Among the new variables added in this chapter, industry and macro-level variables can effectively improve the fraud identification rate of China's online lending platforms, while relevant variables extracted from comments cannot improve the fraud identification rate of China's platforms.
(3) **Compared to individual investors, news commentators in China have a more rational judgment on online lending platforms:** On one hand, the variables extracted from the comments of individual investors on Wangdaizhijia and P2P eye did not improve the recognition rate of platform fraud in China, while the variables extracted from industry news can effectively improve the recognition rate of platform fraud in China. On the other hand, positive news can effectively suppress the possibility of platform fraud, while negative news has a significant promoting effect on platform fraud. Although positive and negative comments have similar effects, their impact is significantly smaller than that of news.
(4) **The direction and magnitude of the effects of different variables on the likelihood of fraud on online lending platforms vary:** Based on the above discussion, it can be concluded that low-risk platforms are characterized by long operating years, having a fund custody mechanism, a bidding guarantee mechanism, relatively high paid-up capital, and being in a good industry and macro environment.

However, the analysis in this section also has the following shortcomings:

(1) **Comment sentiment analysis technology needs improvement:** As discussed above, it can be concluded that variables related to comments are not effective. This conclusion may be related to the irrationality of individual investors in China as well as the imprecise assumptions of sentiment analysis and topic extraction used in this chapter. This chapter aims to reduce the difficulty of training sentiment analysis models. It is assumed that users on P2P Eye have selected sentiment labels that match the sentiment of their comments. However, in reality, there are cases where the labels do not match the sentiment of the comment content. For example, the corresponding label for "platform banking supervision has been launched, and all necessary certificates are available, so it is reassuring!" is "negative review," while the corresponding label for "platform is unreliable and did not honor their promise to offer referral rewards" is a "positive review." This issue may lead to the model learning incorrect patterns, and if processed in a semi-supervised manner, more accurate emotional labels may be obtained.

(2) **Prediction lag:** When constructing features, the processing method for dynamic data is to calculate its mean or proportion during each platform operation period, which covers the entire operation period of the platform. This approach will lead to a lag in predicting the likelihood of fraud on the platform, meaning that once a higher likelihood of fraud is identified, the platform may experience fraud in a short period of time. This lag has little impact on regulators, as regulation is more focused on identifying the variables that will have an impact on platform fraud. If the platform does not have a fund custody mechanism, it is necessary to strengthen monitoring of the platform. However, for investors, the impact of lag is significant, as the value of fraud identification for investors lies in the identification results rather than variables. Bid tenors on existing online lending platforms are concentrated at one month, three months, six months, one year, etc. If investors invest their funds on a certain platform, they need to ensure that the platform will not encounter problems in the shortest possible time after one month so that investors will not suffer losses. However, the model trained in this chapter cannot predict the platform's state beyond a month or longer, which diminishes its value as a reference for investors.

This chapter will further address the drawback of prediction lag in the following section.

8.6 Fraud Detection for Online Loan Platforms

In the previous section, by comparing different models and data combinations, meaningful variables and suitable models for identifying fraud on China's online lending platforms were identified. However, there is a lag in their identification, which can only provide some helpful insight for regulators. For investors, the value of such models is relatively weak. Therefore, this section will adjust the data calculation method in order to predict the platform's status at the end of the investment period and effectively provide early warning for investors.

8.6.1 *Feature construction*

From the analysis in the previous section, it can be concluded that variables related to comments are not effective variables in predicting platform status. Therefore, in the following analysis, seven variables related to comments will be proposed. The remaining variables are classified into two categories: (1) static variables, including 13 variables such as actualCapital, platEarningsCode, autoBidCode, association, bindingFlag, bankFunds, inspection, manageExpense-01, rechargeExpense-01, vipExpense-01, locationCity, newTrustCreditorCode, and platBG; (2) dynamic variables, including eight variables such as onlineTime, plat_num, news_1, news_0, news_-1, risk_free_rate, unemployment_rate, and stock_index_rate.

The calculation method for static variables is the same as in Chapter 4, and an example of calculating dynamic variables is shown in Figure 8.15. The time window is set to three months, which means that the platform's relevant data within three months will be used to provide fraud warnings. The moving step is one month, which means generating a new set of data every other month. The warning period is three months, which means predicting the platform status in three months.

Taking the platform with the launch date of August 20, 2015, and the time when it became a problematic platform of April 30, 2017, as an example, the first set of data is the average of various dynamic data from

August 20, 2015, to November 19, 2015. At this time, the platStatus is 0 and the onlineTime is 1. The second set of data is the average of various dynamic data from September 20, 2015, to December 19, 2015, with a platStatus of 0 and an onlineTime of 2. Similarly, the 16th set of data is the average of various dynamic data from November 20, 2016, to February 19, 2017. At this time, the platStatus is 1 (because the platform had already encountered problems three months later on February 19, 2017), and the onlineTime is 16.

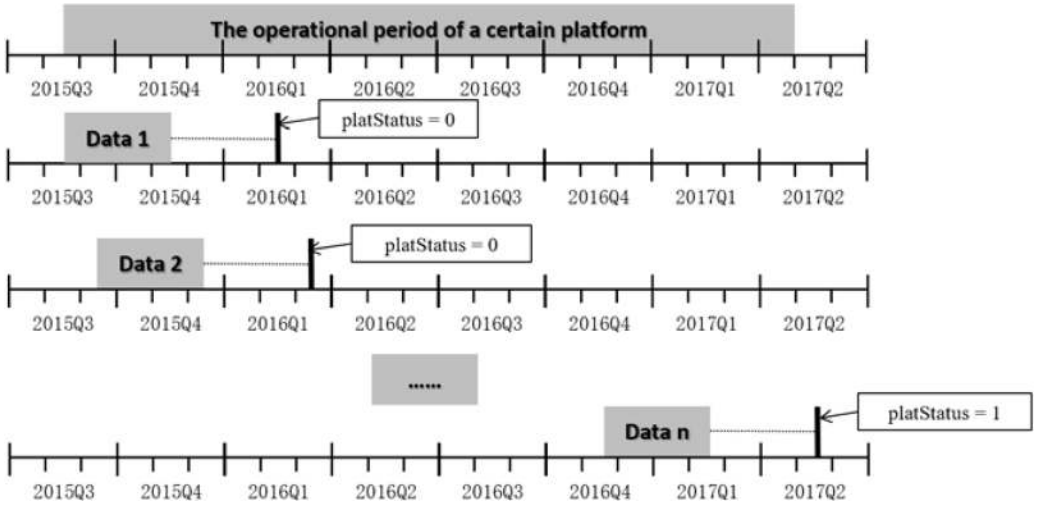


Figure 8.15. Example of dynamic variable calculation.

8.6.2 *Balance treatment*

Through the above processing, 41,763 samples with platStatus 0 and 5,495 samples with platStatus 1 were obtained. There is also a class imbalance problem, but the type of imbalance is different from that in Chapter 4. The imbalance in Chapter 4 is that there are far more positive samples than negative samples, while the imbalance in this section is that there are far more negative samples than positive samples. Since the goal of this chapter is to identify positive examples (i.e., identify platforms that may have issues), the impact of the imbalance types in Section 4 on the model is far less significant than in this section.

In order to address the impact of class imbalance, this section adopts the class balancing processing method based on the Bagging idea proposed by

Liu Hualing *et al*., as shown in Figure 8.16.[99] The training set is divided into datasets with platStatus 1 and platStatus 0, and 3,500 samples are randomly selected from each dataset, combined into a completely balanced new training set with 7,000 data points. The model is trained on the new training set. We repeat this step 19 times to generate 19 models, and finally obtain the result through voting on 19 models.

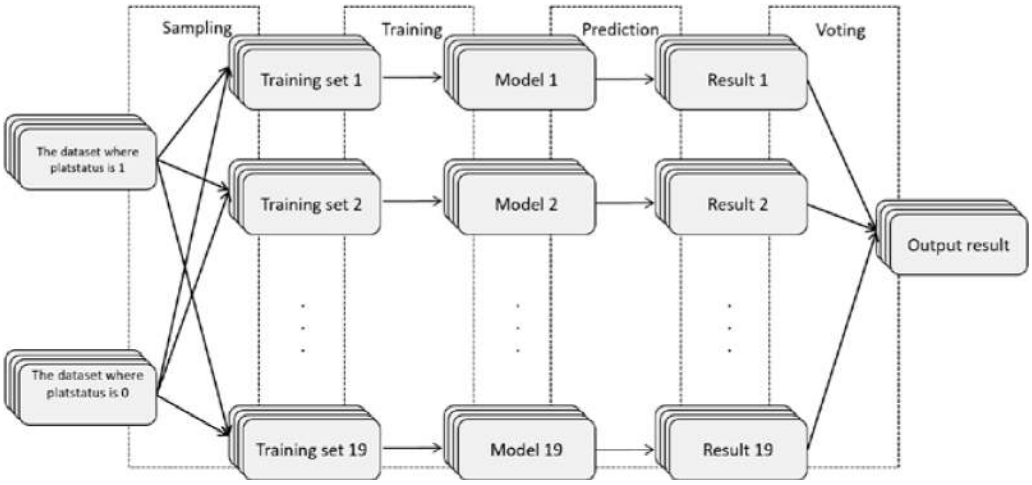


Figure 8.16. Balance processing based on Bagging.

8.6.3 *Fraud detection*

The various indicators of the model trained before and after the above feature construction and balance processing on the training and testing sets are shown in Table 8.5.

8.6.3.1 *Pre balancing*

From the table, it can be seen that the pre-balanced model performs well in AUC and accuracy, and shows average performance in accuracy and F1 Score. However, it shows poor performance in recall, which is only around 0.55 on both the training and testing sets. According to the definition, recall rate measures the proportion of the actual positive samples that are classified as positive, i.e., the probability that a platform where the problem actually occurred is identified. However, 0.55 indicates that only half of the problematic platforms can be identified in advance, which undoubtedly

lacks sufficient value for investors. Once the other half of the platforms are not identified and investors invest in them, they are likely to face losses in both principal and interest. Therefore, although imbalanced samples perform well in some indicators, they perform poorly in the key indicator of recall, making further sample balancing necessary.

Table 8.5. Fraud detection results.

	Pre balancing		Post balancing	
	Training set	**Test set**	**Training set**	**Test set**
AUC	0.94807	0.94640	0.90369	0.89829
Accuracy	0.77669	0.77291	0.86748	0.85253
Precision	0.99718	0.99234	0.55751	0.54515
Recall	0.55359	0.54639	0.82034	0.79277
F1 Score	0.71194	0.70474	0.66385	0.64605

8.6.3.2 *Post balancing*

From the balanced training set results, it can be seen that AUC and F1 Score have decreased compared to those before balancing, but the decrease is not significant. Accuracy and recall have increased compared to those before balancing, and accuracy has decreased significantly compared to that before balancing. In terms of accuracy, the evaluation of the training set decreased from 0.99 before balancing to 0.55 after balancing. According to the definition, accuracy measures the proportion of samples that are actually positive and are classified as positive, i.e., the probability that platforms that have not actually encountered problems are recognized, and 0.55 indicates that nearly half of all platforms that did not encounter problems were misjudged as not potentially problematic. Therefore, from the perspective of accuracy alone, the balanced model did not provide better predictive results. However, in terms of recall rate, the evaluation on the training set has increased from 0.55 before balance to 0.79 after balance, indicating that the balanced model can recognize more problematic platforms. This increase has great significance for investors, greatly reducing the possibility of losses caused by investors investing funds in potential problematic platforms. Therefore, considering both accuracy and recall, this section believes that although the balanced model increases the misjudgment rate

for non-problematic platforms, it reduces the likelihood of losses for investors, making the balanced model more effective.

8.6.3.3 *Summary*

This section adjusts the data calculation method so that the model can predict the likelihood of fraud occurring in three months based on the platform's performance within the past three months. Due to the adjustment of calculation methods, the data is imbalanced (with more negative samples than positive samples), resulting in a significant decrease in the model's recognition rate of the problem platform. In order to improve the recognition rate of problematic platforms and provide better reference for investors, this section carried out a "data+model" balance processing, significantly improving the recognition rate of problematic platforms.

8.7 Conclusion

The rapid development of the Internet has infused new energy into traditional finance. Internet finance has flourished in the past decade with the help of the development of Internet technology. P2P online lending, as an important branch of Internet finance, has entered a rapid development stage after a short accumulation period since it entered China in 2007. However, due to imperfect regulatory measures and irrational investors, problems have occurred frequently on online lending platforms in recent years, with incidents such as platform defaulting without making repayments and "difficulty in withdrawal" occurring frequently. This is undoubtedly unfavorable for China's online lending investors and has a certain impact on the Chinese financial environment.

This phenomenon has attracted the attention of scholars both in China and overseas, and many studies have been conducted. However, existing research mostly focuses on identifying fraud with different bids within a single platform, and there are few scholars conducting fraud identification on different platforms. At the same time, due to the fact that most of the indicators selected by existing research are static indicators, there are few studies that conduct fraud warning based on fraud identification, providing investors with long-term fraud identification.

We have conducted in-depth research in this chapter to address the issues mentioned above and the research conducted in this chapter is characterized by the following:

(1) **Distinct research perspectives:** This chapter adopts a platform-level perspective by comparing the performance of different platforms on the same variable and using a model to determine the possibility of a platform fraud.
(2) **Comprehensive variable selection:** This chapter not only selects commonly used variables in existing research, such as basic platform information, but also incorporates variables such as investor reviews, industry news, and macroeconomic conditions.
(3) **Diverse research objectives:** On the one hand, this chapter identifies fraud and performs variable screening through training models; on the other hand, it refines data calculation methods based on the selected variables to enable fraud warning.

Through data collection, processing, and empirical testing, this chapter draws the following conclusions:

(1) Among the variables added in this chapter, industry-level variables (including platform quantity and industry news) and macro-level variables (including urban unemployment rate and risk-free interest rate) can effectively improve the identification rate of fraud in China's online lending platforms, while variables related to individual investor comments (including comment quantity and emotional inclination) cannot effectively improve the identification rate of fraud on China's online lending platforms.
(2) The Bagging+GBDT fraud warning model constructed in this chapter can effectively improve the identification rate of fraud on China's online lending platforms and provide an effective reference for investors to reduce the possibility of losses caused by choosing the wrong online lending platform.
(3) The fraud warning system constructed in this chapter can provide investors with a more comprehensive overview of the online lending industry and individual platforms.

[a] https://baike.baidu.com/item/%E6%89%98%E7%AE%A1/17704944. ↩

Chapter 9

Internet Financial Risk Assessment Based on Semi-Supervised Model

9.1 Background

The research paradigm of the semi-supervised framework breaks through the limitations of traditional methods that only consider one sample type, which can mine the hidden information of a large amount of unlabeled data to assist with a small number of labeled samples for training. When dealing with credit risk assessment in the presence of few samples, semi-supervised learning frameworks can use the attributes of the data to improve the decision boundaries and make the classification results more accurate than classifiers constructed only using labeled data. Ensemble Learning, on the other hand, runs multiple weak classifiers at the same time when iterating over the training set and filters the classifiers based on certain rules, thus effectively improving the learning effect. Meanwhile, the Ensemble Learning approach has been widely used in the field of financial fraud identification. Research on China's consumer finance market shows that Ensemble Learning methods, such as GBDT and XGBoost, are better than the traditional methods, which is because the data in China's Internet finance market is usually high-dimensional heterogeneous.[100]

Combining the research paradigm of semi-supervised learning with Ensemble Learning tends to have better results when constructing credit risk assessment models. Dietterich's study (1997) points out that the Ensemble Learning method is more effective than the traditional method when assessing the credit risk of a customer.[101] Tumer and Ghosh (1996) use Ensemble Learning to filter and optimize the features extracted based

on the training set, which optimizes the prediction.[102] The research of Zhou and Li (2005) constructively proposed a semi-supervised integration model called Tri-training.[103] The model first learns different base classifiers on different training sets, and if two of the base classifiers classify a sample consistently, this sample will be added to the training set of the third classifier. The above steps will iterate until the preset condition is satisfied and then learning will stop. Finally, the classification results of the three classifiers on the test samples are integrated by the voting method. A subsequent study by them proposed the Co-forest algorithm, which extends the Tri-training algorithm from three classifiers to more classifiers.[104] The collaborative training algorithm obtains the classifiers through multi-view training, and the interaction between the classifiers enables the collaborative training algorithm to improve the accuracy of recognition.[105] The integration of semi-supervised learning and Ensemble Learning methods, which realize the automated labeling of unlabeled samples, provides an effective solution to the missing-label problem in Internet financial credit assessment.

However, existing semi-supervised co-training models do not consider the impact of changes in sample features over time on model performance. Wei Dai *et al.* (2019) proposed an innovative online adaptive semi-supervised learning algorithm to achieve product quality evaluation in industrial engineering, and the research concluded that industrial processes usually have slow time-varying characteristics, and therefore, the established data model should change parameters accordingly.[106] In Internet financial credit risk modeling, the problem of dynamic changes in fraudulent behavior is prevalent: Potential fraudsters are antagonistic to risk assessment, and continuously adapt their strategies to bypass fraud identification models. As a result, static rule-based models struggle to achieve satisfactory results, and the classification mechanism of the risk control model needs to be constantly updated in response to evolving information.

The massive data volume and fast response demands of Internet finance make machine learning algorithms the right choice. Meanwhile, given the abundance of unlabeled samples in credit risk assessment, Ensemble Learning methods based on semi-supervised frameworks represent a promising direction for future research. The key to achieving automated labeling is to utilize a small number of labeled samples to pre-train the

model in a supervised setting. A meta-learning framework is a new research direction for dealing with the task of learning with few samples.[107] The goal of the meta-learning framework is to optimize model parameters by training a meta-learning model on similar learning tasks so that it only needs to use a few training samples.[108] Inspired by the meta-learning framework for financial fraud detection proposed by Abbasi *et al*. (2012)[109] this chapter utilizes multi-rule feature engineering and proposes an adaptive semi-supervised framework for risk assessment model based on three classifiers, GBDT,[110] DART,[111] and GOSS,[112] which can efficiently predict credit defaults in the Internet finance sector.

9.2 Semi-Supervised Learning

9.2.1 *Overview of semi-supervised learning*

In many practical applications of machine learning, it is easy to find a large number of unlabeled samples, but it is necessary to manually label them using special equipment or through an expensive and very long experimental process in order to obtain labeled samples, which results in a very small number of labeled samples and a surplus of unlabeled samples. Therefore, an attempt was made to learn by adding a large number of unlabeled samples to the limited number of labeled samples to be trained together, with the expectation that it would improve the learning performance, thus giving rise to semi-supervised learning.

9.2.2 *Assumptions of semi-supervised learning*

The establishment of semi-supervised learning relies on modeling assumptions, and when the modeling assumptions are correct, unlabeled samples can help improve learning performance. Semi-supervised learning relies on the following three assumptions:

(1) **Smoothness Assumption:** Two samples located at a short distance have similar class labels in a dense data region. When two samples are connected by an edge in a dense data region, there is a high probability

that they have the same class label. Conversely, when two samples are separated by a sparse data region, their class labels tend to be different.

(2) **Cluster Assumption:** When two samples are in the same cluster, there is high probability that they have the same class label. The equivalent definition of this assumption is the assumption of Low Density Separation (LDS), which states that the classification decision boundary should pass through the sparse data region and avoid cutting through the dense data region, so that similar samples are not separated by the decision boundary.

(3) **Manifold Assumption:** When high-dimensional data are embedded into a low-dimensional manifold, two samples are likely to share the same class label if they are located in a small local neighborhood within the manifold. Many experimental studies have shown that when SSL does not satisfy these assumptions or when the model assumptions are incorrect, the samples without class labels not only fail to improve the learning performance but also deteriorate it, leading to a degradation of SSL's performance. However, some other experiments have shown that unlabeled samples have the potential to impair learning performance even if the model assumptions are correct in some unusual cases.

9.2.3 *Classification of semi-supervised learning*

According to the statistical learning theory, semi-supervised learning can be categorized into two types of modes: transductive semi-supervised learning and inductive semi-supervised learning. Transductive semi-supervised learning only deals with the given training data in the sample space: it utilizes both labeled and unlabeled samples in the training data to predict the class labels of the unlabeled samples in the training data. Inductive semi-supervised learning processes both the training data and unseen samples in the entire sample space. It uses labeled and unlabeled training samples together with unseen test samples, not only to predict the labels of unlabeled training samples but also, more importantly, to predict the class labels of unlabeled test samples.

Looking at different learning scenarios, semi-supervised learning can be categorized into four main groups:

(1) **Semi-supervised classification:** With the help of samples without class labels, the training samples with class labels obtain a classifier with better performance than the one obtained by training only with samples with class labels, compensating for the lack of labeled samples, where the class labels y_i take finite discrete values $y_i \in \{c_1, c_2, \ldots, c_c\}$, $c_j \in N$.

(2) **Semi-supervised regression:** With the help of inputs without outputs, training the inputs with outputs obtains a regressor that performs better than the regressor obtained by training only the inputs with outputs, where the outputs y_i take continuous values and $y_i \in R$.

(3) **Semi-supervised clustering:** With the help of information from labeled samples, this method obtains better clusters than the results obtained with only unlabeled samples, improving the accuracy of clustering methods.

(4) **Semi-supervised dimensionality reduction:** With the help of information from the samples with class labels, find the low-dimensional structure of the high-dimensional input data while keeping the structure of the original high-dimensional data and Pair-Wise Constraints unchanged. That is to say, samples that satisfy Must-Link Constraints in the high-dimensional space are very close to each other in the low-dimensional space, and samples that satisfy Cannot-Link Constraints in the high-dimensional space are very far away from each other in the low-dimensional space.

9.3 Research Design

9.3.1 *General idea of credit risk assessment model*

The general idea of the credit risk assessment model based on an adaptive semi-supervised framework constructed in this chapter is shown in Figure 9.1. In this chapter, the preprocessed dataset is divided into a labeled training set, an unlabeled training set, and a test set. The basic idea is to first train multiple different classifiers using a labeled training set and traditional supervised learning methods and then expand the training set with each classifier's predictions on unlabeled data. The key issues in this process are

how to construct multiple different classifiers and which unlabeled data to choose to expand the training set.

The construction of a credit risk assessment model based on an adaptive semi-supervisory framework mainly includes the following steps:

(1) **Training phase:** Based on the data in the labeled training set *L*, soft-feature and hard-feature are extracted by setting the evaluation criteria and applying the LightGBM algorithm. Four training subsets are constructed by different sampling ratios, in which the ratios of non-overdue data to overdue data in the first three subsets are 3:1, 5:1, and 7:1, respectively, and the last subset contains all data. The base classifiers are trained separately in the above subsets with different sampling ratios. Using a supervised learning approach, the training process produces the Stack Model List.

(2) **Labeling phase:** Adaptive learning is done on the unlabeled dataset *U* using the models in the model list, and if the predicted labels generated by the model list for the unlabeled samples do not match, then the unlabeled samples are put back without labeling. If the predicted labels are consistent, the sample is considered to have passed the consistency test, and the sample is labeled accordingly, the process is shown in Figure 9.2. The samples that passed the consistency test will have their features compressed and extracted using the CNN model and then be put back into the model of the training set for further training.

(3) **Test phase:** The samples in the test set *T* are finally classified to obtain the predicted labels in the test set.

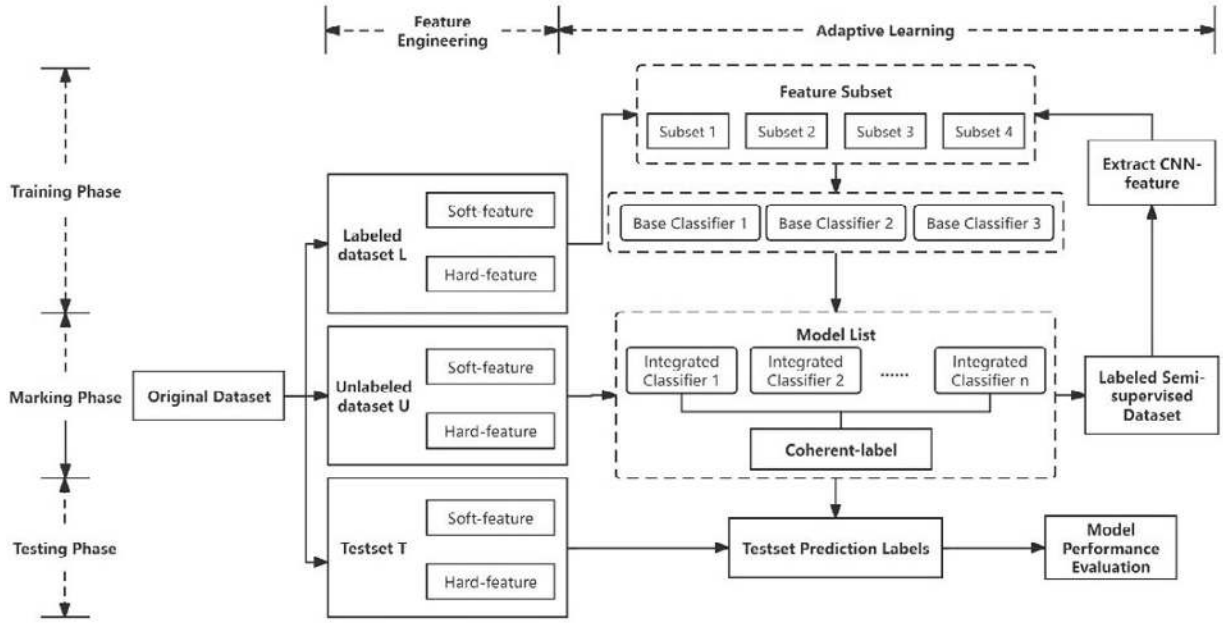


Figure 9.1. General idea of the credit risk assessment model.

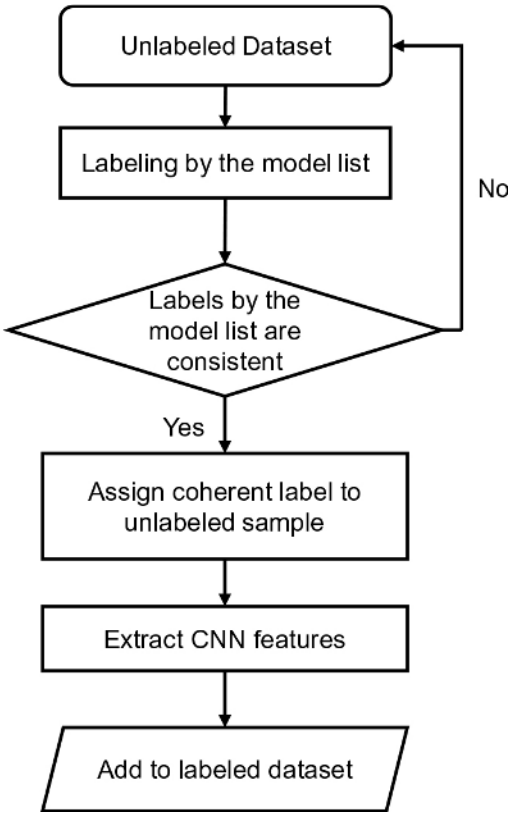


Figure 9.2. Labeling process.

9.3.2 *Adaptive learning framework*

Given the inherent adversarial relationship between potential fraudsters and credit assessment systems, financial fraud recognition is a dynamic process. Therefore, the performance of static rule-based classifiers will gradually deteriorate, and the classification method needs to be constantly revised with updated information.[109] Adaptive learning methods are particularly relevant in the context of Internet financial risk assessment; they can be aware of the changes in their risk environment by learning new data and can constantly retrain themselves accordingly, making the credit risk detection system more adaptive to the new situation, so as to identify potential fraudsters trying to camouflage their behavior and detect new fraud patterns.

Inspired by the research on meta-learning frameworks for financial fraud detection, this chapter constructs a similar adaptive learning framework, as shown in Figure 9.3. A model list including 12 models was

generated through training based on the four feature subsets of a labeled training set and three models, GBDT, DART, and GOSS. The 12 models in the model list are then used to adaptively learn from the unlabeled dataset, and if all 12 models agree that the sample is either overdue or not, the sample is considered to have passed the consistency test and will be labeled with the appropriate coherent label, which ensures that only samples with strong predictive consistency are added to the training data. After adding this labeled data to the training set, the model is retrained, and the remaining data in the unlabeled dataset are then labeled. This resets the training set data during each iteration and reclassifies all training set samples to correct errors and strengthen the model list for credit risk identification while enabling automatic labeling of unlabeled data.

The reason for constructing the feature subset is that traditional co-training methods can only train classifiers in a single view, and in this study, different views of the data are used for learning to get a differentiated model, which in turn leads to the construction of a model list. The classifiers obtained from training in different subsets can be sensitive to distinctive features, thus achieving complementary advantages. The unlabeled data is then fed into the model list for prediction, and the predictions are finally integrated to obtain the class label for that data. Due to the differences in the feature subsets, the classifiers in the former model list are differentiated. At this point, if all the classifiers are still able to generate a coherent label, then the label is extremely credible. In other words, a coherent label generated by differentiated classifiers can reduce the labeling error.

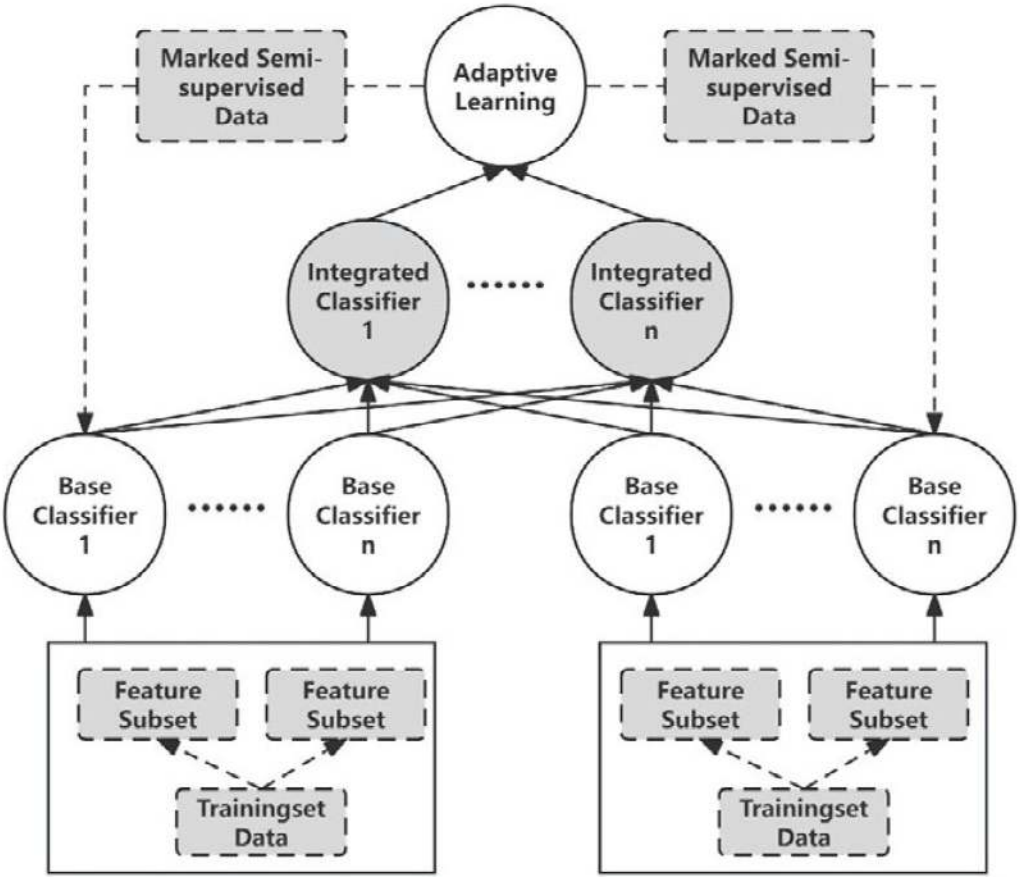


Figure 9.3. The adaptive learning framework.

9.3.3 *Feature engineering in credit risk assessment*

The feature engineering constructed in this chapter consists of two parts. In the first part, based on the training set data, soft-feature and hard-feature are extracted by setting evaluation criteria and applying the LightGBM algorithm, respectively. In the second part, in order to fully utilize all the features, a CNN model is constructed for feature extraction.

9.3.3.1 *Soft-feature*

When data preprocessing is complete, the features need to be selected for input into the machine learning algorithms and models for training. Any missing values or zeros in the data need to be eliminated. For data with a very small range of variation, the variation of the column feature is captured

by defining an informativeness metric (info = mean(x)/std(x)), and data with a small degree of variation also need to be removed.

The filter selects features by selecting the full feature vector space from the training data and filtering and searching within this space to get the features within the threshold and determine the final feature set. The selection evaluation criterion used in this chapter is the Spearman rank correlation coefficient, and the data are graded and filtered according to the threshold value of the correlation coefficient.

9.3.3.2 *Hard-feature*

LightGBM algorithm is applied to the data in a training set with category labeling $\boldsymbol{L}$ for supervised learning, and the output feature set is used as hard-feature. LightGBM adopts the histogram algorithm, which is based on the idea of discretizing the continuous floating-point features into $\boldsymbol{k}$ discrete values and constructing the histogram with width $\boldsymbol{k}$. Then, it will traverse the training data and calculate the cumulative statistics of each discrete value in the histogram. In feature selection, we only need to traverse the histogram to find the optimal partition point based on the discrete values of the histogram.

9.3.3.3 *CNN-feature*

A large amount of data in the training set is eliminated when hard-feature and soft-feature are used for the feature screening process. To fully utilize all the features, a CNN-feature extraction model is constructed in this chapter. All the data learned through adaptive learning constitute a new dataset called Labeled Sime-Data, based on which features are extracted using a CNN for compression.

The basic structure of the convolutional neural network constructed in this chapter to extract features consists of four layers, with the first layer being the feature input layer. The second and third layers are feature extraction layers. Each neuron in this layer is connected to the local receptive domain of the previous layer, and the convolution kernel is used to extract the local features. Once the local feature is extracted, the positional relationship between it and other features is also determined, and the feature mapping structure adopts the ReLU function as the activation function of the convolutional network in order to avoid the disappearance of

the gradient. Each convolutional layer in the convolutional neural network is immediately followed by a Max-Pool pooling layer used for feature compression and secondary extraction, and this secondary feature extraction structure reduces the feature dimension. The resulting output layer includes 64 features.

9.3.4 *Adaptive semi-supervised algorithmic process*

The flow of the adaptive semi-supervised algorithm constructed in this chapter is shown in Figure 9.4, where the labeled training set $D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, and unlabeled training set $D_m = \{x_l, x_{l+1}, \ldots, x_{l+m}\}$ are taken as inputs, and the outputs are the classification results of the final trained classification model $H_T = \{h_1(x), h_2(x), \ldots, h_T(x)\}$.

The training subset D_l is all labeled data, and then three training subsets, D_{l1}, D_{l2}, and D_{l3}, are constructed by sampling, and the ratios of non-overdue data to overdue data are 3:1, 5:1, and 7:1, respectively. Three classifiers, M_1, M_2, and M_3, are trained on D_{l1}, D_{l2}, and D_{l3}, respectively, and the list of models containing 12 classifiers is obtained as $C = \{C1, C2, \ldots, C_T\}$.

(1) Put each data in the unlabeled training set D_m into the model list $C = \{C_1, C_2, \ldots, C_T\}$ for consistency test, and let the classification result obtained by the ith unlabeled training sample on the tth classifier be Z_{it}:

$$Z_{it} = h_t(l_{t+i}),\ i \in (1, m),\ t \in (1, T) \tag{9.1}$$

(2) Let the label prediction of the ith unlabeled training sample be $\hat{y}_{l+i}$:

$$y_{l+i} = \begin{cases} 1 & \sum_{t=1}^{T} Z_{it} = T \\ 0 & \sum_{t=1}^{T} Z_{it} = 0 \\ None\ \ else \end{cases} \tag{9.2}$$

If all models in the model list consistently label the unlabeled training sample as 1, it is $\sum_{t=1}^{T} Z_{it} = T$, then the Coherent-label of the unlabeled training sample is assigned as 1 by the Coherent-label test; if all models in the model list label the unlabeled training sample as 0, it is $\sum_{t=1}^{T} Z_{it} = 0$, then the Coherent-label of the unlabeled training sample is assigned as 0 by the Coherent-label test; in other cases, nothing is done. The label of the unlabeled training sample is assigned as 0; in other cases, nothing is done.

(3) Add the sample $(x_{l+i}, \hat{y}_{l+i})$ with $\hat{y}_{l+i} = 0$ or 1 to the labeled semi-supervised dataset D_{semi}.

(4) The labeled semi-supervised dataset D_{semi} and the labeled training set D_{l1} are merged to get a new training set $D_{l'}$, which is then put into the model list $C = \{C_1, C_2, \ldots, C_T\}$ for training to get the output results.

Input: Training set with category labels $D_{l1} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$;
Unlabeled training set $D_m = \{x_l, x_{l+1}, \cdots, x_{l+m}\}$;
Base classifiers M_1、M_2、M_3;

Process:

1. Random sampling to obtain feature subsets D_{l2}、D_{l3}、D_{l4};
2. Three classifiers M_1、M_2、M_3 are trained separately on D_{l1}、D_{l2}、D_{l3}、D_{l4} to obtain a list of models $C = \{C_1, C_2, \cdots, C_T\}$;
3. **for** $t = 1,2,\cdots,T$ **do**
   $h_t = C_t(D_m)$
   end for
4. $D_{semi} = \phi$;
5. **for** $i = 1,2,\cdots,m$ **do**
   for $t = 1,2,\cdots,T$ **do**
   $Z_{it} = h_t(x_{l+i})$;
   end for
   end for
6. $$\hat{y}_{l+i} = \begin{cases} 1, & \sum_{t=1}^{T} Z_{it} = T \\ 0, & \sum_{t=1}^{T} Z_{it} = 0 \\ None, & else \end{cases}$$
7. **for** $i = 1,2,\cdots,m$ **do**
   if $\hat{y}_{l+1} = 0$ **or** $\hat{y}_{l+1} = 1$
   $D_{semi} = D_{semi} \cup \{(x_{l+i}, y_{l+i})\}$;
   end for
8. $D_l' = D_{l1} \cup D_{semi}$;
9. **for** $t = 1,2,\cdots,T$ **do**
10. $h_t = C_t(D_l')$;

Output: $H_T = \{h_1(x), h_2(x), \cdots h_T(x)\}$

Figure 9.4. Adaptive semi-supervised algorithm flow.

9.4 Simulation Experiment

9.4.1 *Experimental dataset*

The experiments in this chapter are based on data such as characteristics and credit scores of a sample of users who lent money from January 1, 2018, to May 1, 2018, on an online lending platform, and the dataset contains a total of more than 100,000 pieces of data, which need to be

preprocessed because the original dataset contains a large number of missing values and zero values.

Of the experimental dataset, about 30% have labels showing whether the sample is overdue while it is unknown whether the remaining 70% is overdue or not. We divided the 30% labeled samples into a labeled training set and a test set in a ratio of 7:3. The unlabeled dataset contributes to 70% of the overall dataset and it is unknown whether these users are overdue or not.

9.4.2 *Model performance evaluation criteria*

To evaluate the performance of the customer credit assessment model, this study uses the AUC evaluation metric. The customer credit evaluation matrix is first introduced (Table 9.1). TP represents the number of correctly predicted positive categories; FN represents the number of incorrectly predicted negative categories; FP refers to the number of incorrectly predicted positive categories; and TN represents the number of correctly predicted negative categories.

Table 9.1. Customer credit evaluation matrix.

	Predicted to be positive category	**Predicted to be negative category**
Actual positive class (overdue clients)	TP	FN
Actual negative class (not overdue clients)	FP	TN

In the binary classification problem, the false positve rate is defined as FPR = FP/(FP + TN), the true positive rate is defined as TPR = TP/(TP + FN), and the ROC curve takes FPR as the *X*-axis and TPR as the *Y*-axis, and the AUC value is the area of the portion of the ROC curve surrounded by the coordinate axises. Since the ROC curve often cannot clearly reflect the performance of the classifier in practice, using the AUC value as the model performance evaluation can better quantify the model's ability to distinguish positive and negative classes. The larger AUC value, the better the discriminative ability of the model.

In order to analyze the impact of the adaptive semi-supervisory framework proposed in this chapter on the performance of the credit scoring model, this chapter constructs a model performance enhancement metric: It is obtained by dividing the difference between the AUC values of the model using the adaptive semi-supervisory framework model and the model under the non-semi-supervisory framework by the AUC value of the model under the non-semi-supervisory framework, and the formula is as follows:

$$\text{Lift} = \frac{\text{AUC}_{\text{ASL}} - \text{AUC}_{\text{NASL}}}{\text{AUC}_{\text{NASL}}} \tag{9.3}$$

where Lift is the model performance improvement degree, AUC_{ASL} is the AUC value of the model using the adaptive semi-supervised framework, and AUC_{NASL} is the AUC value of the model in the non-semi-supervised framework.

9.4.3 *Experimental results analysis*

9.4.3.1 *Impact of CNN-feature on model performance*

Table 9.2 shows the experimental results of the three models, GBDT, DART, and GOSS, learned in the adaptive semi-supervised framework on training sets with different proportions of class distributions and with or without the incorporation of CNN-features, respectively. For example, in the first and fourth columns, the AUC value obtained by the GBDT model with the inclusion of CNN-features trained on the dataset D_{l2} is 0.8460, while the AUC value of the model without CNN-features is only 0.8404.

Comparing the performance of the models with and without the inclusion of CNN-features, it can be found that all three classifiers containing CNN-features have higher AUC values than the model without CNN-features trained on the feature subset, but the classification performance on the full set is slightly worse. The results in Table 9.2 show that the inclusion of CNN-features improves the model prediction performance in the adaptive semi-supervised framework.

Table 9.2. Model performance in the semi-supervised framework.

Dataset	Adding CNN-features	Without adding CNN-features

	GBDT	DART	GOSS	GBDT	DART	GOSS
D_{l1}	0.8419	0.8458	0.8430	0.8497	0.8369	0.8469
D_{l2}	0.8460	0.8458	0.8489	0.8404	0.8307	0.8404
D_{l3}	0.8408	0.8393	0.8394	0.8391	0.8364	0.8373
D_{l4}	0.8477	0.8471	0.8470	0.8384	0.8366	0.8402

Table 9.3. Model performance in the unsupervised framework.

	Adding CNN-features			Without adding CNN-features		
Dataset	**GBDT**	**DART**	**GOSS**	**GBDT**	**DART**	**GOSS**
D_{l1}	0.8416	0.8437	0.8484	0.8310	0.8237	0.8292
D_{l2}	0.8448	0.8455	0.8369	0.8400	0.8378	0.8455
D_{l3}	0.8469	0.8534	0.8399	0.8397	0.8363	0.8308
D_{l4}	0.8449	0.8503	0.8497	0.8462	0.8428	0.8324

9.4.3.2 *Impact of an adaptive semi-supervised framework on model performance*

Table 9.3 shows the experimental results of the three models in the unsupervised framework with different proportions of class distributions in the training set and with or without the inclusion of CNN-features.

Comparing the results in Tables 9.2 and 9.3, it can be found that the AUC values of the three classification models under the adaptive semi-supervised framework are larger than those under the non-semi-supervised framework, regardless of whether CNN-features are included or not. Therefore, the adaptive semi-supervised framework constructed in this chapter can significantly improve the model performance and realize automatic labeling for unlabeled training samples, which provides a new solution to the problem of large amounts of unlabeled customer data in credit assessment in Internet finance.

In Figure 9.5, the y-coordinate represents Lift, which was defined earlier in this chapter as the degree of performance improvement achieved by the adaptive semi-supervised framework. This metric is obtained by dividing the difference in AUC values between the model using the adaptive semi-supervised framework model and the model under the non-semi-supervised framework by the AUC value of the model under the

non-semi-supervised framework. The x-coordinate denotes the training data subsets for the three models — GBDT, DART, and GOSS — characterized by different class proportions and by whether CNN-features are included. A Lift greater than zero indicates that the use of the semi-supervised framework can improve the model performance. The results show that the adaptive semi-supervised framework constructed in this chapter has the most obvious improvement effect on the performance of GOSS models. Overall, the semi-supervised co-training can effectively improve the credit assessment performance of the three models without adding CNN-features on the training set that includes all the features; the use of the adaptive semi-supervised framework with the addition of CNN-features on a subset of features with different class proportions can also improve the prediction quality of the GBDT and DART models to a certain extent.

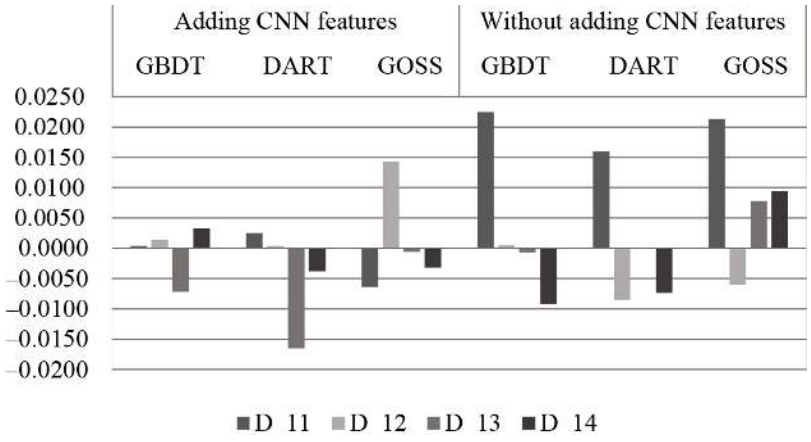


Figure 9.5. Impact of adaptive semi-supervised framework on model effectiveness.

9.5 Conclusion

In this chapter, a credit risk assessment model for Internet finance is designed based on a semi-supervised framework using adaptive learning. This model employs an innovative fusion of multiple methods for feature extraction. Based on four training subsets and three Ensemble Learning models, GBDT, DART, and GOSS, a list of differentiated models containing 12 integrated classifiers is trained. The predictions generated by

each classifier on unlabeled data are then used to expand the training set of the classification models, thus avoiding wasting most of the effective data. The results of simulation experiments show that the combination of feature engineering and random sampling for generating feature subsets, as proposed in this chapter, has significantly improved the effectiveness of the model. Overall, the adaptive semi-supervised framework has better credit assessment performance and can realize automatic labeling for unlabeled training samples, which on the one hand provides a new solution to the problem of large amounts of unlabeled customer data in credit assessment and on the other hand provides an empirical basis for the selection of credit assessment models in Internet finance.

Chapter 10

A Study of Risk Measurement of the Stock Index Futures Market Based on Quantile Regression Forests

10.1 Background

The rapid development of the economy and information technology has led to financial innovations and the emergence of a variety of financial derivative products, which provide more investment options for investors in the financial market. Stock index futures, as one of the most important financial derivatives, has the inherent attributes of high leverage and high risk, and investments in stock index futures products are subject to a greater number and intensity of risk factors as compared to vanilla financial assets. Therefore, it is crucially important to portray the risk of the stock index futures market accurately, and it is imperative to adopt scientific and reasonable methods to control and manage such risks.

In his study of stock index futures risk, Engle (1982) proposed the Autoregressive Conditional Heteroskedasticity (ARCH) model, which could address the issues of volatility clustering, leptokurtosis, and heavy tails that occur in financial time series.[113] Subsequently, his student Bolerslev (1986) proposed a GARCH model based on the ARCH model by taking the influence of volatility into consideration.[114] Afterward, a series of extended models such as EGARCH and GARCH-M were developed based on the GARCH model (Nelson,[116] Engle *et al.*,[117] Rabemananjara and Zakoian[118]). Currently, the GARCH-VaR model has become the most common risk measurement method in finance. Li and Xue (2013)

conducted an empirical analysis of the GARCH-VaR model using the trading data of the CSI 300 stock index futures and argued that the model proves to be an effective risk management tool for investors to a certain extent.[115] Yu and Wang (2018) estimated the VaR of the CSI 300 Index based on a proposed Mix Frequency Realized GARCH model (M-Realized GARCH model).[119] Ma Ying *et al.* (2019) analyzed the VaR risk of international oil prices using the SWARCH model and compared it with the ARCH family model, and the results of the study showed that the prediction accuracy of the SWARCH model is higher than that of the ARCH family model for measuring the volatility of international oil price.[120] However, GARCH models usually have to make assumptions about the distribution of the return series first, which tends to affect the forecasting results of VaR to some extent.

The quantile regression method was first proposed by Koenker and Bassett (1978), which became a major breakthrough in the field of statistics because it was possible to characterize the effect of the explanatory variable on the response variable at different quantiles without assuming the distribution of returns series.[121] Many scholars have conducted further research on VaR using the quantile regression method. Li Bangguo (2017) studied the risk of the soybean futures market using quantile regression, and the results showed an increase in the accuracy of VaR values compared to classical regression models.[122] Bayer (2018) introduced “a regu-larization parameter” in quantile regression.[123] Liu Yansen (2017) proposed a Bayesian Lasso quantile regression and used it for VaR value estimation in financial markets.[124]

Quantile regression effectively captures the impact of explanatory variables on the response variable at different quantiles in linear problems. In practice, however, the relationship between explanatory and response variables may be nonlinear. Quantile regression forest is a derivative algorithm of Breiman's random forest proposed by Meinshausen and Ridgeway (2006), and for regression problems, random forests provide not only the conditional mean of the dependent variable but also the complete information on its conditional distribution.[125] The quantile regression forest method can effectively fit the complex nonlinear structure that may exist in the sequence. Park *et al.* (2018) used quantile regression forest to predict monthly extremes for the next 20 years based on historical precipitation

data and proved the effectiveness of the method by testing on real data.[126] Sun Guoqiang *et al.* (2018) proposed to apply quantile regression forests to probability density prediction of short-term wind power. The results showed that compared with the traditional short-term wind power prediction methods, quantile regression forests are able to provide predictive results under any quantile point, so as to obtain the conditional distribution of predicted values, and the probability density predictive results of short-term wind power are more accurate than before.[127]

This chapter examines the VaR of stock index futures using both the quantile regression forest algorithm and traditional quantile regression, and compares these approaches with traditional VaR measures. The results show that the quantile regression forest algorithm achieves higher measurement accuracy compared to traditional models.

10.2 Model Theory

10.2.1 *GARCH model*

Assuming that the daily return rate of the financial time series is r, the general time series model of r can be expressed as follows:

$$r_t = \varphi_0 + \sum_{i=1}^{m} \varphi_i r_{t-i} + a_i - \sum_{j=1}^{n} \phi_j a_{t-j}, \quad a_t = \sigma_t \varepsilon_t \tag{10.1}$$

$$\sigma_t^2 = \widetilde{\omega} + \sum_{i=1}^{p} a_i a_{t-i}^2 + \sum_{j=1}^{q} \gamma_j \sigma_{t-j}^2 \tag{10.2}$$

Equation (10.1) is the conditional mean process for y_t, under the ARMA(m, n) model, usually assuming that ε_t follows a standard normal distribution; Eq. (10.2) is the conditional variance process for y_t, following the GARCH(p, q) model.

On the basis of Eqs. (10.1) and (10.2), further calculations yield a one-step forward projection of the conditional mean $\widehat{r_t}(1)$ and a one-step forward projection of the conditional variance $\sigma_t^2(1)$ of the return series y_t as follows:

$$\widehat{r_t}(1) = \varphi_0 + \sum_{i=1}^{m} \varphi_i r_{t+1-i} - \sum_{j=1}^{n} \phi_j a_{t+1-i} \tag{10.3}$$

$$\sigma_t^2(1) = \sigma + \sum_{i=1}^{p} a_i a_{t+1-i}^2 - \sum_{j=1}^{q} \gamma_j \sigma_{t+1-i}^2 \tag{10.4}$$

Assuming that all historical information Γ_{t-1} is available at moment t, y_{t+1} then follows a conditional normal distribution $y_\{t+1\}|\Gamma_t$ with conditional mean $\widehat{y_t}\left(1\right)$ and conditional variance $\sigma_t^2\left(1\right)$, and follows a normal distribution $N\left(\widehat{y_t}\left(1\right), \widehat{\sigma_t^2}\left(1\right)\right)$. Let τ represent the upper tail probability; the VaR for a holding period of 1 day can be computed at moment t as follows:

$$\mathrm{VaR}_t(1-\tau) = -y_t(1) - \sigma_t(1) z(\tau) \tag{10.5}$$

In Eq. (10.5), $z\left(\tau\right) = \Gamma_z^{-1}\left(\tau\right)$ represents the quantile of the standard normal distribution.

10.2.2 *Quantile regression*

Building on regression to the mean, quantile regression constructs a model for a specific quantile of interest by using data corresponding to that quantile of the explanatory and response variables. This approach allows one to examine the effect of the explanatory variable on the response variable at any given quantile, thereby providing insight into the full conditional distribution of the response variable, the information at any quantile, and the relationship between the response and explanatory variables.

Assuming the random variable $F(Y)$ = prob($Y \leq y$) has the following probability distribution function

$$F(Y) = \text{prob}(Y \leq y) \tag{10.6}$$

for $0 < \theta < 1$, the quantile of Y can be defined as follows:

$$Q(\theta) = \inf \{y : F(y) \geq \theta\} \tag{10.7}$$

For financial time series $\{y_t, x_t\}_{t=1}^{n}$, y_t is the rate of return on financial assets and x_t is an explanatory variable that affects the return on financial assets. Then, the conditional quantile of y_t can be expressed as follows:

$$Q_\nu(\tau) = \inf\left\{y \in R : F(y|x) \geq \tau\right\} = F^{-1}(\tau|x) \tag{10.8}$$

Also, it is equivalent to solving the following optimization problem:

$$\min \sum^{n} E[(\tau - I\left(y_t < Q_v\left(\tau\right)\right)\left(y_t - Q_v\left(\tau\right)\right)|x_t = x] \tag{10.9}$$

where $I(\bullet)$ is the indicator function, as follows:

$$I\left(\omega\right) = \begin{cases} 0, \, \omega \geq 0 \\ 1, \, \omega < 0 \end{cases} \tag{10.10}$$

VaR is defined as the maximum loss that an investor can suffer by holding a particular financial asset or portfolio of assets for a period of time at a confidence level of $100(1 - \tau)\%$. Assuming that the cumulative distribution function of the loss L of a financial asset or portfolio of assets is y_t, the VaR at a confidence level of $100(1 - \tau)\%$ is the quantile of the distribution of the rate of return, that is,

$$\mathrm{VaR}_{y_t} = -Q_{y_t}(\tau) \tag{10.11}$$

The VaR under the quantile regression model can thus be obtained.

10.1.1 *Quantile regression forest*

The linear quantile regression method proposed by Koenker *et al.* reflects well the effect of the explanatory variable on the response variable at different quantiles in a linear problem. In practice, however, the relationship between explanatory and response variables may be nonlinear, and the quantile regression forest model proposed by Meinshausen and Ridgway (2006) can effectively fit the complex nonlinear structure that may exist in the series. The quantile regression forest builds on quantile regression by introducing random forests to simulate the nonlinear part of the return series.

The conditional distribution function for Y given $X = x$ is as follows:

$$F(y|X=x) = P(Y \leq y|X=x) = E(I_{\{Y \leq y\}}|X=x) \tag{10.12}$$

We define the estimate of the last term of the expression $E(I_{\{Y \leq y\}}|X = x)$ to be the weighted average of the observations $I_{\{Y \leq y\}}$ as follows:

$$\widehat{F}(y\Big|X=x) = \sum_{i=1}^{n} \omega_i(x) I_{\{Y \leq y\}} \tag{10.13}$$

where the weight of ω_t is the mean of the decision tree. The estimation of $\widehat{F}(y\Big|X=x)$ can be executed using the following algorithm:

For new data, k decision trees $T(\theta_t)$, $t = 1, \ldots, k$ are first generated. The first step is to get the predicted values for each tree by considering the observations on each leaf node of each tree, and then voting or weighted averaging all the predicted values to get the prediction for a single tree. Similarly, when constructing a random forest, the predicted values of all decision trees, which should be first considered, are voted or weighted averaged to get the final prediction of the random forest. To perform a weighted average, the weights of the leaf nodes and the weights of the single decision tree are of importance. The formula for calculating the weights of each leaf node of each decision tree is given first. Suppose an observation X_i belongs to a leaf node (x, θ) and is not 0; let the weight vector $\omega_i(x, \theta)$ be

$$\omega_i(x, \theta) = \frac{1\{X_i \in R_l(x, \theta)\}}{\#\{j : X_j \in R_l(x, \theta)\}}. \tag{10.14}$$

The sum of weights $\omega_i(x, \theta)$ is equal to 1. The weights of the leaf nodes can then be used to obtain the predicted values for each decision tree:

$$\widehat{u}(x) = \sum_{i=1}^{n} \omega_i(x, \theta) Y_i. \tag{10.15}$$

After obtaining the predicted values from each tree, the conditional mean under the random forest, $E(Y|X = x)$, is computed as a weighted average of each decision tree's predictions. Assume that the random vectors $\theta_t(t = 1, \ldots, k)$ of each decision tree are independently and identically distributed. Let $\omega_i(x)$ represent the mean of the k decision trees, as follows:

$$\omega_i(x) = k^{-1} \sum_{t=1}^{k} \omega_i(x, \theta_t) y. \tag{10.16}$$

For all $y \in R$, on the basis of $\omega_i(x)$, the estimate $\widehat{Q_\alpha}\left(x\right)$ of the conditional quantile $Q_\alpha(x)$ of the distribution function can be replaced by $\widehat{Q_\alpha}\left(x\right) = \inf\left\{y : \widehat{F}\left(y\middle|X = x\right) \geq \alpha\right\}$, where $\widehat{F}\left(y\middle|X = x\right)$ is calculated by the previous algorithm.

10.3 Empirical Research

10.3.1 *Data selection*

The data selected for the research and analysis in this chapter, spanning from January 4, 2011 to September 28, 2018, were taken from the CSI 300 stock index futures. The data were downloaded to get the daily closing data of the SSE Composite Index from the iFinD software, and there are 1,884 samples in total. The stock index returns were converted to log returns via a natural logarithmic transformation. Where the logarithmic return formula is $r_{i,k} = \text{In}(p_{i,k}) - \text{In}(p_t)$ and k is the holding period. In this chapter, the holding period determined in all constructed models is 1 day, and the returns data are logarithmic daily returns, that is, $r_t = \text{In}(p_{t+1}) - \text{In}(p_t)$.

10.3.2 *Correlation statistics test*

Figure 10.1 shows the histogram of the frequency distribution of the CSI 300 stock index futures returns. The curve is the density distribution curve fitted with a normal distribution. From the figure, we can initially determine that the distribution of returns does not follow a normal distribution, but is

characterized by sharp peaks and thick tails. Further, we test the normality of the CSI 300 stock index futures. Both the QQ plot of the distribution of returns in Figure 10.2 and the results of the KS test in Table 10.1 reject the null hypothesis that the distribution of index returns follows a normal distribution.

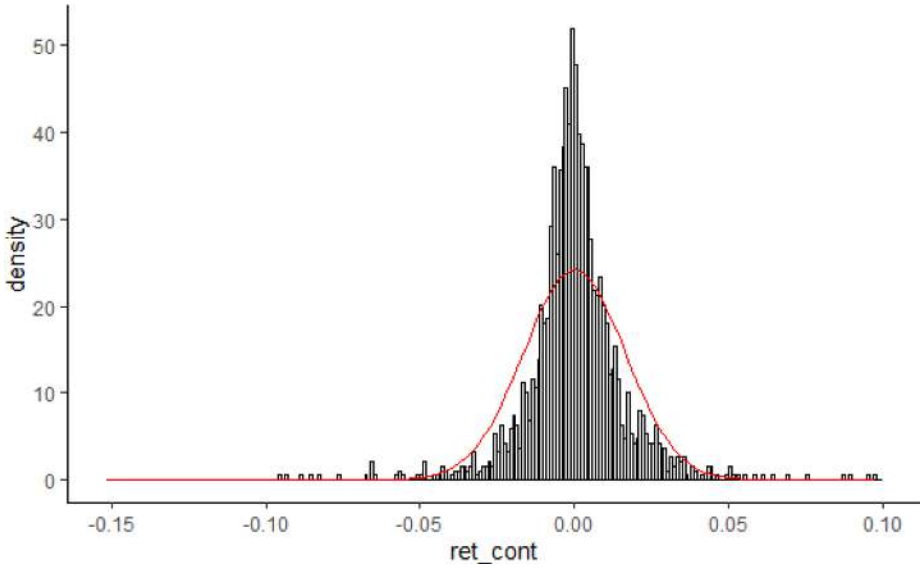


Figure 10.1. Frequency histogram of returns on CSI 300 stock index futures.

Figure 10.3 shows a chart of the logarithmic return of the CSI 300 index futures, in which the return fluctuates around the mean value. Figure 10.4 shows an autocorrelation plot of logarithmic returns, in which the returns on the CSI 300 stock index futures converge to 0. Therefore, we tentatively conclude that the return series of the CSI 300 stock index futures is smooth. Further, we test the unit root of the return series based on the ADF test, PP test, and KPSS test, where the null hypothesis of the ADF test and PP test is that the series has a unit root, while the null hypothesis of the KPSS test is that the series is smooth. The results of all three unit root tests in Table 10.2 indicate that the return series of the CSI 300 stock index futures is smooth.

Figure 10.4 shows a clear volatility clustering in the return series of the CSI 300 stock index futures. We test the ARCH effect on the returns, and the Ljung–Box test in Table 10.3 rejects the null hypothesis that the volatility series is not autocorrelated, suggesting that there is an ARCH effect on the return series of the CSI 300 stock index futures.

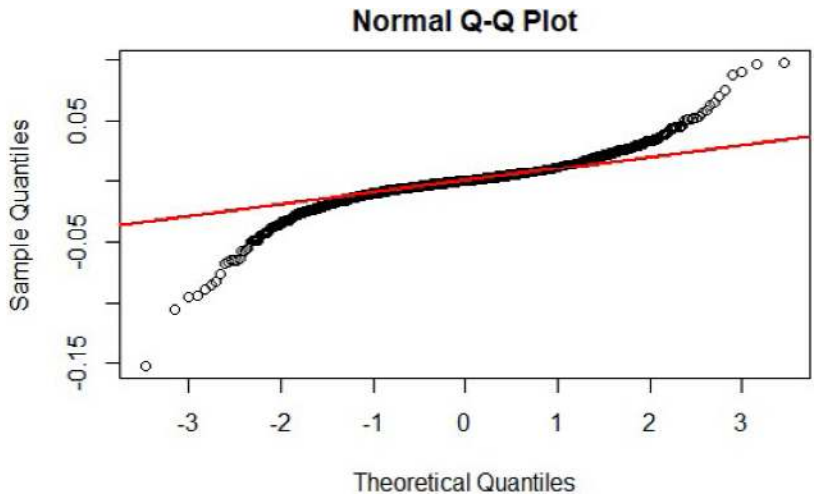


Figure 10.2. QQ chart of CSI 300 stock index futures returns distribution.

Table 10.1. KS test for logarithmic returns of CSI 300 stock index futures.

	D	***P-value***
KS test	490.46	2.2e-16

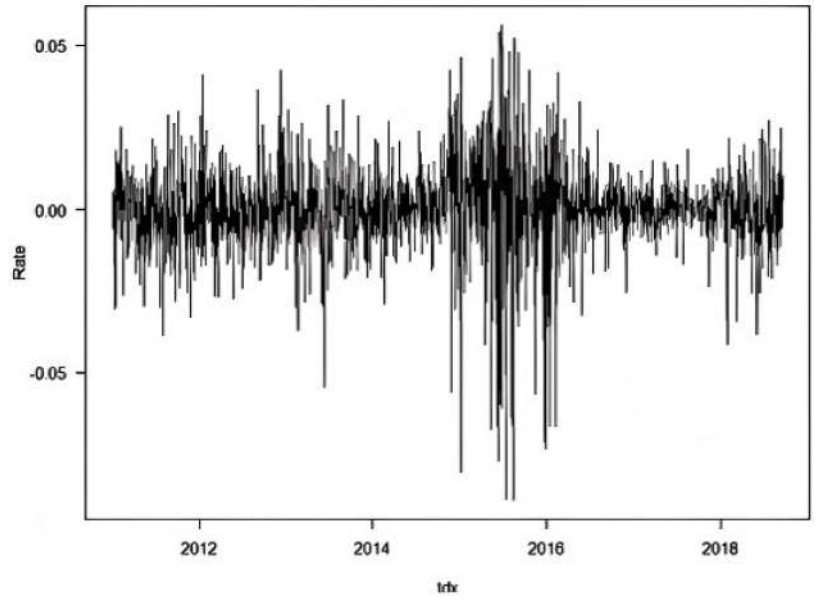


Figure 10.3. CSI 300 stock index futures log yield frequency chart.

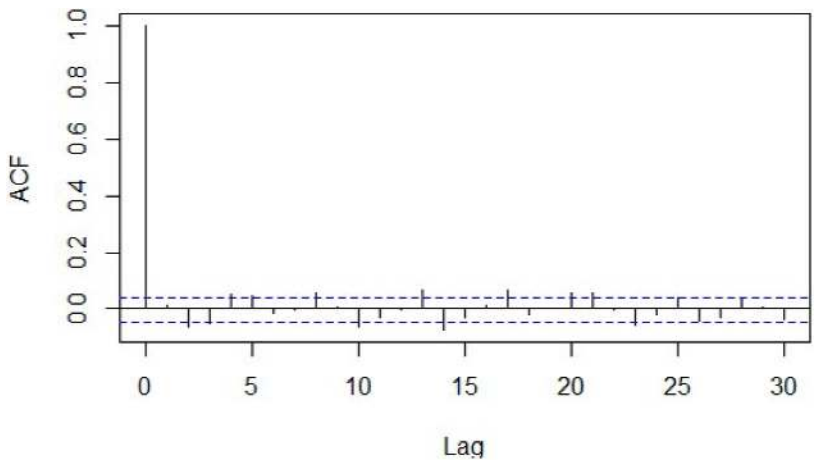


Figure 10.4. Autocorrelation coefficient of CSI 300 stock index futures returns.

Table 10.2. Unit root test of CSI 300 stock index futures returns.

	T-statistic	***P*-value**
ADF test	−11.919	0.01
PP test	−42.625	0.01
KPSS test	0.1027	

Note: For the KPSS test, the critical values at 10%, 5%, 2.5%, and 1%, respectively, were 0.347, 0.463, 0.574, and 0.739.

Table 10.3. ARCH effect test of CSI 300 stock index futures returns.

	Chi-square value	***P-value***
Ljung–Box test	490.46	2.2e-16

In conclusion, the CSI 300 stock index futures return series is a smooth series, and the distribution shows sharp peaks and thick tails, asymmetry, and conditional heteroskedasticity.

10.3.3 *Model building*

10.3.3.1 *VaR measurement based on GARCH models*

The statistical characteristics of the CSI 300 stock index futures return series indicate that there is an ARCH effect in the return series, so a GARCH model can be built to capture the changes in its volatility. A high-level summary of the VaR-GARCH model is to establish an optimal GARCH model to analyze the historical return series of the Shanghai Stock Exchange Composite Index (SSEI), get the standard deviation of the GARCH prediction through its variance equation, assume that the SSEI obeys the *T*-distribution, select the corresponding probability density standard value, and accurately compute the value of VaR for each trading day. All data processing and modeling in this chapter were carried out using the statistical software R. The quantile regression model was established and parameter estimation was carried out. The results of parameter estimation of the GARCH model at the significance levels of $\tau = 0.01$, 0.05, and 0.1 are shown in Table 10.4.

After obtaining conditional volatility under the GARCH model, the VaR under the GARCH model is calculated and compared with the original return series. The VaR series and the return series are superimposed on a graph, as shown in Figure 10.5.

Figure 10.5 shows that the higher the confidence level, the better the performance of VaR, and the more the fluctuations in returns that can be covered; regardless of whether the confidence level is 99%, 95%, or 90%, the change in VaR is in the same direction as that of the actual data, and the VaR curve is able to encompass most of the fluctuations in returns.

Table 10.4. Parameter estimates for the GARCH model.

	α_1	**1.1.1**	γ_1	**1.1.2**
GARCH(1,1)	0.05	1.5e−13	0.95	2.2e−16

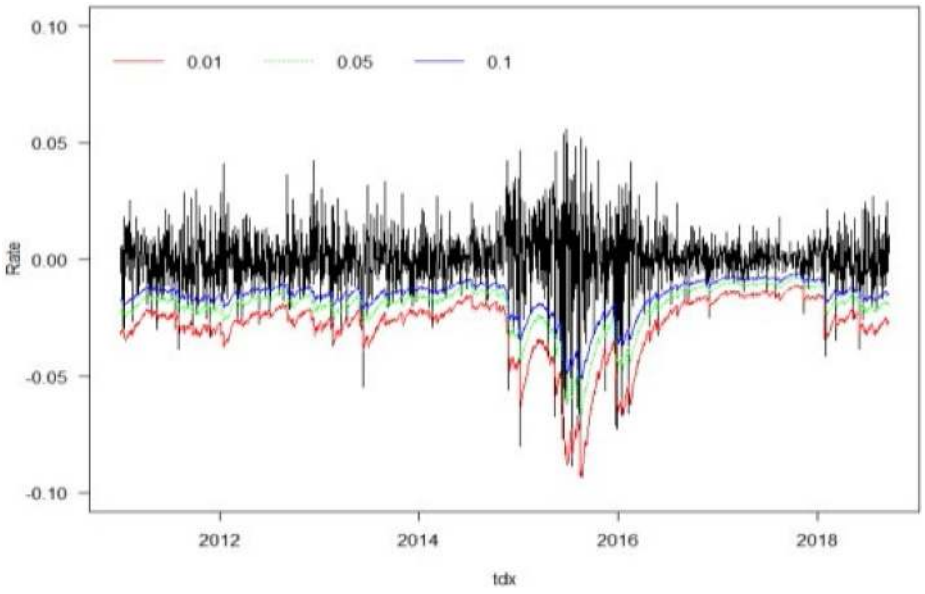


Figure 10.5. Plot of log returns and VaR sequences under the GARCH model.

10.3.3.2 *VaR measurement based on QR models*

In this section, we mainly use the quantile regression model proposed by Chen and Chen (2002) to conduct the VaR risk measurement study, and the model structure is as follows[128]:

$$r_t = \beta_0 + \beta_1 \sigma_t + \beta_2 \sigma_t^2 + \varepsilon_t \tag{10.17}$$

where σ_t is estimated using the GARCH(1,1) model.

The following points can be noted from Table 10.5: (1) In general, the overall significance of the model is not significant and the parameters are less significant at significance levels of 0.05 and 0.1 than at a significance level of 0.01. (2) The higher the confidence level, the better the VaR model performs. (3) The coefficient of the quadratic term is insignificant at either confidence level, that is, the linear quantile regression model fails to capture its nonlinear structure.

From Figure 10.6, it can be seen that regardless of whether the confidence level is 99%, 95%, or 90%, the movement of the VaR series and the movement of the actual return series are consistent. The VaR curve is able to cover the vast majority of the volatility of the return. Moreover,

consistent with the above-mentioned analysis, the higher the confidence level, the more the return volatility can be covered by the VaR.

Table 10.5. QR_GARCH model parameter estimates.

Significance level	99%	95%	90%
β_0	0.01689	0.00057	−0.00206
$p\beta_0$	0.07167	0.84813	0.56919
β_1	−5.42041	−1.67526	−0.86283
$p\beta_1$	0.00005	0.00309	0.01504
β_2	65.46605	−2.96858	−6.81028
$p\beta_2$	0.06248	0.8892	0.76214

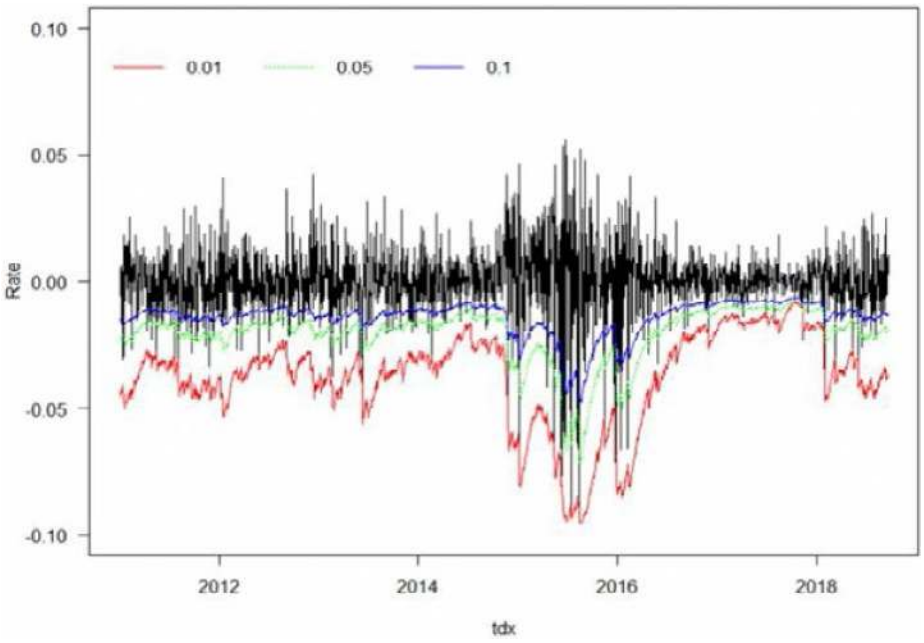


Figure 10.6. Quantile regression series of CSI 300 stock index futures returns and VaR.

10.3.3.3 *VaR measurement based on QRF models*

The quantile regression forest is a derivation of random forest, which is mainly based on quantile regression and introduces the random forest to capture the nonlinear structure of the financial system. It not only provides full information on the conditional distribution of the dependent variable

but also improves the accuracy of conditional quantile estimation. Given the advantages of the quantile regression algorithm and the lack of prior applications of the quantile regression forest algorithm in academic research on risk measurement, this chapter is the first attempt to use the quantile regression forest algorithm as a new approach for VaR calculation. Quantile regression forest does not simply take a quantile regression approach to minimize the loss function, but is based on the random forest approach. The model employs features drawn from those used by Chen in the quantile regression model described earlier. The implementation of the model is based on the quantregForest package of the *R* language, and the model uses the default parameters, with the specific values shown in Table 10.6.

The following quantile regression forest model is established to obtain the VaR estimates at different confidence levels respectively, and the VaR with quantile points of 0.01, 0.05, 0.1 respectively are compared with the CSI 300 stock index futures return series, and the VaR series and the return overlay are shown in Figure 10.7.

From Figures 10.5–10.7, the following results can be obtained: (1) A vast majority of the VaR values under the three models can cover the volatility of returns. (2) The higher the confidence level, the higher the VaR inclusion. (3) The VaR curve of the quantile regression forest is more inclusive. However, it is impossible to judge the performance of the model only from the graphs; the Kupiec likelihood ratio test is utilized next to compare the VaR values obtained from the above-mentioned model predictions with the actual loss data and to test the accuracy and validity of the models by quantifying the gap between them.

Table 10.6. Parameters used in random forest modeling.

ntree	ntry	Nodesize
500	43	5

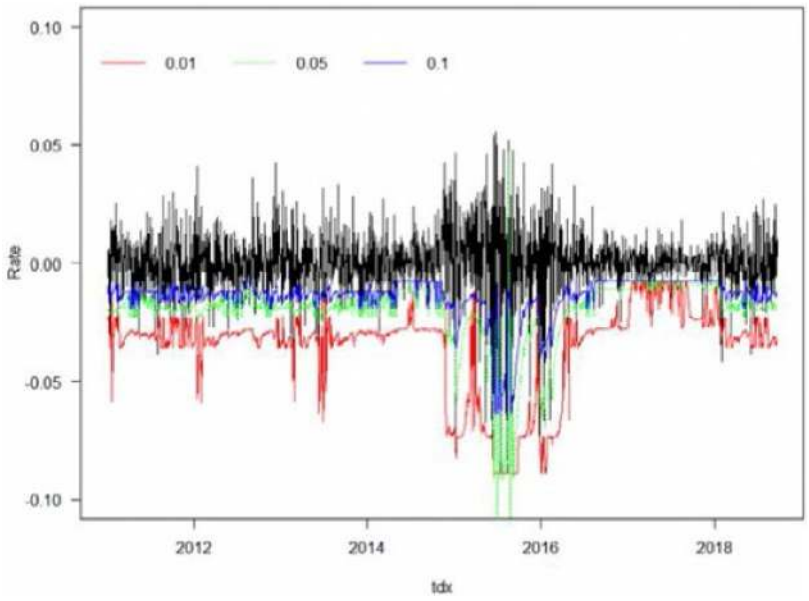


Figure 10.7. Quartile regression forest of CSI 300 stock index futures log returns and VaR sequences.

10.3.4 *Comparative analysis results*

In order to accurately capture the validity of the estimated results of the VaR model, it is necessary to test the accuracy of the VaR. A commonly used test is the Kupiec likelihood ratio test.[149] This method focuses on testing whether the estimates are in line with reality. It defines the percentage of times when the actual gains and losses in the observed sample exceed the VaR estimates as the failure rate. It assumes that the number of failure days N within a sample observation period of T days follows a binomial distribution (T, p), with p being the failure rate. The null hypothesis of the Kupiec test is $\alpha = p$, and the statistic of its likelihood ratio test

$$\mathrm{LR} = -2\ln\left[(1-\alpha)^{N-N_0}\alpha^{N_0}\right] + 2l \tag{10.18}$$

Under the condition that the null hypothesis holds, the likelihood ratio statistic will follow a χ^2 (1) distribution.

In the following, the Kupiec likelihood ratio test is utilized to quantify the gap between the predicted value of VaR and observed actual losses, We

then compare the quantile regression forest model, quantile regression model, and GARCH model to determine which model can more accurately predict the VaR of the Shanghai Composite Index. The selected sample includes 1884 pieces of data, and the data are from the iFinD software. First of all, by comparing the obtained VaR series with the actual loss series, the number of days of failure can be obtained, which in turn can be calculated to obtain the failure rate; then, we use the Kupiec likelihood ratio test to test the failure rate and select the significance level of 0.01, 0.05, and 0.1, which can be calculated to obtain the LR statistic for different models with different confidence levels. The results obtained are shown in Table 10.7.

The following results are obtained from Table 10.7: (1) At confidence levels of 99%, 95%, and 90%, the LR statistics of quantile regression forest are 0.0375044, 0.008054, and 0.002888, respectively; the LR statistics of quantile regression are 0.243513, 0.014843, and 0.02888, respectively; and those of the GARCH model are 19.73061, 0.512979, and 7.35576, respectively. The LR statistics of the quantile regression forest are smaller than those of quantile regression and the GARCH model, and the results show that the quantile regression forest model performs better. (2) At the confidence levels of 99%, 95%, and 90%, as observed from Table 10.7, the gap between the failure rate of the quantile regression forest and the confidence level is also smaller than that of quantile regression and the GARCH model. Thus, the quantile regression forest model performs better than others in VaR prediction. The quantile regression forest is able to capture the complex structure of financial time series better than others, and is thus more suitable for analysis of financial data.

Table 10.7. Results of the Kupiec likelihood ratio test.

Model	Confidence level	Days of failure	Failure rate	LR
GARCH	99%	41	2.18%	19.731
	95%	101	5.25%	0.513
	90%	154	8.33%	7.356
	99%	21	1.11%	0.244
Quantile regression	95%	92	4.88%	0.015
	90%	185	9.82%	0.029
Quantile regression forests	99%	18	0.96%	0.038
	95%	95	5.04%	0.008
	90%	189	10.03%	0.003

10.4 Conclusion

In this chapter, the historical returns of the CSI 300 stock index futures are modeled and their parameters are estimated using three models, namely, GARCH, quantile regression, and quantile regression forest. Furthermore, comparative studies are conducted. The empirical results show that relative to the traditional linear model, the quantile regression forest model can accurately and consistently give an estimate of the VaR risk measure; it can effectively capture the nonlinear relationship of the financial time series and fit the complex nonlinear structure that may exist in the financial time series. Moreover, the VaR model based on the quantile regression forest can better reflect the market risk of stock index futures.

Chapter 11

Research on VaR Risk Measurement Based on the EMD-QRF Model

11.1 Introduction

Safeguarding financial risks is important for maintaining the bottom line of systemic financial risks and improving risk management and early warning mechanisms. Faced with increasingly diverse, complex, and uncertain financial markets, researching effective methods to measure financial risks can enhance the accuracy of risk prediction, facilitating financial institutions and investors in formulating corresponding strategies.

Value at risk (VaR) of financial assets is an important risk measurement indicator. It defines the maximum loss that an asset may suffer when held by investors at a given confidence level under normal market conditions. Currently, VaR measurement methods can be divided into three types: parametric, semi-parametric, and non-parametric methods. Parametric methods assume that daily returns follow a certain distribution, leading to parameter fitting and subsequently deriving VaR based on the statistical properties of the distribution. The Variance–Covariance method developed by JP Morgan is the earliest parametric method proposed, which assumes that asset returns follow a normal distribution and requires the variance to remain constant. Puneet *et al*. (2021) extended the parametric method of VaR estimation by applying transformations for skewness and kurtosis, providing an alternative approach based on the Variance–Covariance method.[129] As the assumption of normal distribution is difficult to meet when studying returns time series in practical research, and financial data generally exhibit characteristics of leptokurtic and heavy tails, scholars gradually relax the assumption of normal distribution, opting for skewed *t*-distribution, generalized error distribution, etc. Bollerslev proposed a novel VaR estimation method based on the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model in 1986, which has significant advantages in characterizing heteroscedasticity and volatility clustering of returns.[114] With the accelerated development of GARCH models in the financial field, a series of studies combining traditional GARCH models have emerged. Youssef (2015) evaluated the risk value and expected loss of crude oil and natural gas commodities markets using three long-memory models, FIGARCH,

HYGARCH, and FIAPARCH, demonstrating that FIGARCH could better predict risk values.[130] Wozniak (2015) utilized constant conditional correlation GARCH family of models to model risk variables in financial time series and to test for Granger causality between second conditional moments, indicating that this method can relax the high-order moment assumption required in classical tests.[131] Wang *et al*. (2017) constructed a novel BEMD-Copula-GARCH model for more effective risk measurement of portfolios, which showed significantly better estimation accuracy compared to DCC-GARCH and Copula-GARCH models.[132] Valizadeh *et al*. (2021) proposed a new HGARCH model with time-varying amplitude and discussed the moment characteristics of the model.[133] They also conducted VaR tests to evaluate predictability, with simulation and empirical results providing further support for the proposed model. In general, it is difficult to accurately obtain the distribution of risk and return by using parametric models, and even if the distribution is obtained, it is difficult to guarantee that its parameters and form will remain unchanged with market changes, thus giving rise to semi-parametric methods for measuring VaR. Semi-parametric methods mainly include tail estimation methods based on extreme value theory and first four moments method. Fan and Gu (2003) estimated the volatility of investment portfolio market prices using semi-parametric methods and proposed both parametric and non-parametric techniques to estimate quantiles of standardized return processes.[134] The results showed that this technique is flexible and can adapt automatically to dynamic market shifts. Non-parametric methods estimate VaR by analyzing existing historical data without making any assumptions about returns, thereby addressing the limitations of parametric methods. Non-parametric methods mainly include kernel density estimation, historical simulation method,[135] quantile estimation, and Monte Carlo simulation method,[136] among which quantile estimation based on quantile regression (QR) is widely used. In terms of statistical mathematics, VaR is essentially a specific quantile of the profit and loss distribution of financial assets. Koenker and others (1978) innovatively proposed a QR model without assuming the distribution of return sequences, which can characterize the influence of explanatory variables on the response variables at different quantiles, offering a method for VaR risk measurement.[137] Wang Xinyu and Zhao Shaojuan (2008) empirically studied the risk of the Shanghai and Shenzhen stock markets using conditional QR models with three different types of explanatory variables, showing that QR is suitable for VaR estimation of financial time series at high confidence levels.[138] Sebastian Bayer (2018)[139] introduced regularization terms in QR to estimate VaR, and Liu Yansen (2017)[140] proposed Bayesian Lasso QR, both showing the accuracy of QR in VaR estimation.

With the increasing demand for accuracy in VaR, people have chosen to combine traditional VaR methods with machine learning algorithms for risk measurement. Xu Qifa *et al*. (2014) proposed that neural network QR methods help improve the accuracy of VaR risk measurement both in-sample and out-of-sample.[141] In the same year, Xu Qifa, Zhang Jinxiu, and others used support vector QR to realize multi-period VaR

measurement, showing that this model can estimate VaR well compared to traditional VaR methods.[142] Sun Wenfeng and Wang Feixing (2017) combined the Least Squares Support Vector Machine with the traditional Monte Carlo simulation method to estimate the volatility of the Shanghai Composite Index.[143] In recent years, Fang Kuangnan (2012) first used the quantile regression forest (QRF) method to estimate VaR for the next day, showing good estimation results of this method.[144] QRF is a derivative algorithm of Breiman's random forest, which was proposed by Meinshausen and Ridgeway.[125] For regression problems, random forests can not only provide conditional means of dependent variables but also provide full information of conditional distribution of dependent variables. The method of using QRF can effectively fit complex non-linear structures that may exist in sequences. Gou Xiaoju and Wang Qian (2019) applied the QRF algorithm to estimate the VaR of the S&P 500 and Shanghai Composite Index and evaluated the impact of weights of various factors on risk values, further validating the effectiveness of this method.[145]

Financial markets are influenced by various factors, such as the economy, national policies, investor psychology, and sentiment. The characteristics, scope, and duration of financial asset price fluctuations caused by each factor are also different. Denoising financial time series data and extracting valuable information are of great significance for predicting financial data. Empirical Mode Decomposition (EMD) is an adaptive time-frequency signal processing method proposed by Huang and others (1998) based on the data's own time scale characteristics for signal decomposition.[146] EMD has higher decomposition efficiency and better time-frequency characteristics, and it has obvious advantages in processing and analyzing various non-linear and non-stationary time series, especially in predicting financial risks and handling unstructured information in financial data. At the same time, the CSI 300 Index is a well-known index in China, with large trading volume, strong liquidity and broad market representation. It is the core index to measure the development of A shares in China. Therefore, researching the VaR measurement of the CSI 300 Index holds important practical implications. This chapter proposes a VaR estimation method based on the combination of EMD and QRF, measures the VaR of the CSI 300 Index, and uses the Kupiec backtesting method to empirically compare this estimation method with various QR VaR methods. The results show that the EMD-QRF model has higher econometric accuracy and generalization ability, providing an effective measurement method for financial risk.

11.2 Research Method

11.2.1 *EMD decomposition*

EMD is an adaptive time series decomposition technique, essentially a process of stabilizing non-stationary data. The EMD method decomposes the fluctuations or trends of different signals in the original signal, producing a series of components with low correlation and different frequencies, each of which is an Intrinsic Mode Function (IMF). Each IMF component reflects the local characteristics of the original signal at different scales. IMF has two constraints: (1) the number of extreme points and zero crossing points in the original signal must be equal or differ by at most one; (2) at any moment, the upper and lower envelopes are locally symmetric about the time axis.

The basic process of the EMD algorithm is as follows:

(1) Identify all local maxima and minima points of the sequence ($R(t)$).
(2) Construct upper envelope ($S(t)$) and lower envelope ($X(t)$) using cubic spline interpolation for local maxima and minima points.
(3) Calculate the mean of $S(t)$ and $X(t)$ to obtain the mean envelope ($M_{11}(t)$) and the middle signal ($H_{11}(t) = R(t) - M_{11}(t)$).
(4) Determine if the middle signal satisfies the two conditions of IMF. If so, the signal is an IMF component; if not, repeat steps 1–4 with this signal as the basis.
(5) After obtaining the first IMF using the above method, subtract it from the original signal ($I_1(t) = R(t) - \mathrm{IMF}_1$), which will become the new original signal. Repeat the above four steps to obtain the second IMF, and so on, until the final component meets the stopping criteria for filtering. Generally, the EMD decomposition of the original sequence ends when the residual component ($I_k(t)$) is less than a certain specified value or becomes a monotonic function. The remaining component $I_k(t)$ represents the trend item or residue of the original sequence.

The decomposition form obtained is as follows:

$$S\left(t\right) = \sum_{i=0}^{k} \mathrm{IMF}_i\left(t\right) + I_k\left(t\right) \tag{11.1}$$

The above equation satisfies the identity relation, implying that the EMD algorithm can decompose the original sequence into multiple components without signal loss, retaining all information of the original sequence.

11.2.2 *Basic principles of quantile regression*

Quantile regression estimates the relationships between explanatory variables X and the conditional quantiles of the response variable Y at all quantile levels.

Under given conditions $X = [x_1, x_2, \ldots, x_k]$, the conditional distribution function is the cumulative probability of $Y \leq y$, which is expressed as

$$F(y_t|X_t) = P(Y \leq (y_t|X_t)) \tag{11.2}$$

The quantile $Q_\alpha(X_t)$ is such that under the given conditions $X_t = [x_{1t}, x_{2t}, \ldots, x_{kt}]$, the cumulative probability of the conditional distribution function of $Y(Y \geq Q_\alpha(X))$ is exactly α, which is expressed as

$$Q_\alpha(X_t) = \inf\{y: F(y_t|X_t) \geq \alpha\} \tag{11.3}$$

where $\inf\{\bullet\}$ is the minimization operation.

The general form of QR is expressed as

$$Q_{yt}(\alpha|X_t) = f_\alpha(X_t, \beta(\alpha)) = \beta_\alpha{}^T X_t \tag{11.4}$$

where β_α is the regression coefficient vector, whose value changes with α. β_α can be solved by the following formula:

$$\underset{\beta}{\operatorname{argmin}} \sum_{t=0}^{T} \rho_\alpha(y_t - \beta_\alpha X_t) \tag{11.5}$$

where T is the sample size, and the test function ρ_α satisfies

$$\rho_\alpha = \begin{cases} \alpha\mu, & \mu > 0 \\ (\alpha - 1)\mu, & \mu < 0 \end{cases} \tag{11.6}$$

Here,

$$\mu = 2\ \operatorname{sgn}(y_t - \beta(\alpha)X_t) - 1 \tag{11.7}$$

VaR is defined as the maximum loss that an investor may suffer when holding specific financial assets for a period at a given confidence level $1 - \alpha$. If the distribution function of the financial assets or portfolio losses L is $F_L(\mathrm{l})$, then the VaR at the confidence level $1 - \alpha$ is the $(1 - \alpha)$th quantile of the return distribution, i.e.,

$$\mathrm{VaR}_{y_t} = Q_{y_t}(\alpha) \tag{11.8}$$

11.2.3 *Quantile regression forest*

QRF is a derivative algorithm of the random forest algorithm. The key difference between QRF and Random Forest is as follows: For each node in each tree, Random Forest retains only the average of the observed values falling on that node and ignores all other information. In contrast, QRF retains all observed values in the node, not just their averages and, based on this, obtains the conditional distribution of the response

variable. For each $X = [x_1, x_2, \ldots, x_k]$, a weighted set of n original observed values $w_i(X)(i = 1, 2, \ldots, n)$ can be obtained. Generally, the QRF algorithm includes the following steps:

(1) Generate k decision trees and examine all observation values at each node of each decision tree.
(2) For a given X, traverse all decision trees, calculate weights $w_i(X,\theta)$, and then take the average of the weights ($w_i(X,\theta)$) to obtain the weight for each observation $w_i(X)$, where

$$w_i(X,\theta) = \frac{1\{X_i \in R_1(x,\theta)\}}{\#\{j : X_i \in R_1(x,\theta)\}} \tag{11.9}$$

$$w_i(X) = \frac{\sum_{t=1}^{k} w_1(x,\theta)Y_1}{k} \tag{11.10}$$

(3) Utilize the obtained weights to estimate the conditional distribution, where

$$F(y|X) = \sum_{i=1}^{n} w_1(x)1(Y \le y) \tag{11.11}$$

(4) The conditional quantile is

$$Q_\alpha(X_t) = \inf\{y: F(y_t|X_t) \ge \alpha\} \tag{11.12}$$

11.2.4 *Kupiec Likelihood Ratio (LR) Backtesting Method*

The most commonly used statistical test method for VaR model estimation results is the LR failure rate test proposed by Kupiec (1995).[149] Specifically, N_0 is the number of instances observed in the series where actual returns exceeding the VaRs estimated by the model. N is the total number of observations in the series. Then the failure rate will be $p = N_0/N$ and its likelihood ratio statistic is:

$$LR = -2l + 2l \sim \chi^2(1) \tag{11.13}$$

11.3 VaR Risk Measurement Based on EMD-QRF

In response to the characteristics of financial time series data — namely, high randomness and signal-to-noise ratio — this chapter combines the advantages of EMD and QRF to establish a VaR measurement model based on the EMD-QRF model. The EMD algorithm is used to decompose the original time series into high-frequency components, low-frequency components, and trend components. The high-frequency component reflects the random fluctuations of the return series and can be characterized using the QRF algorithm. The low-frequency component and trend component are relatively stable in the short term and can be fitted with short-term historical data. Finally, the estimation results of the three components are superimposed to obtain the VaR value of the original series. Figure 11.1 shows the flowchart of estimating the VaR of financial assets using the EMD-QRF model. First, the EMD algorithm is used to process the original time series, extract the market's different frequency components, and establish predictive models accordingly, thereby effectively improving the prediction accuracy of financial assets' VaR.

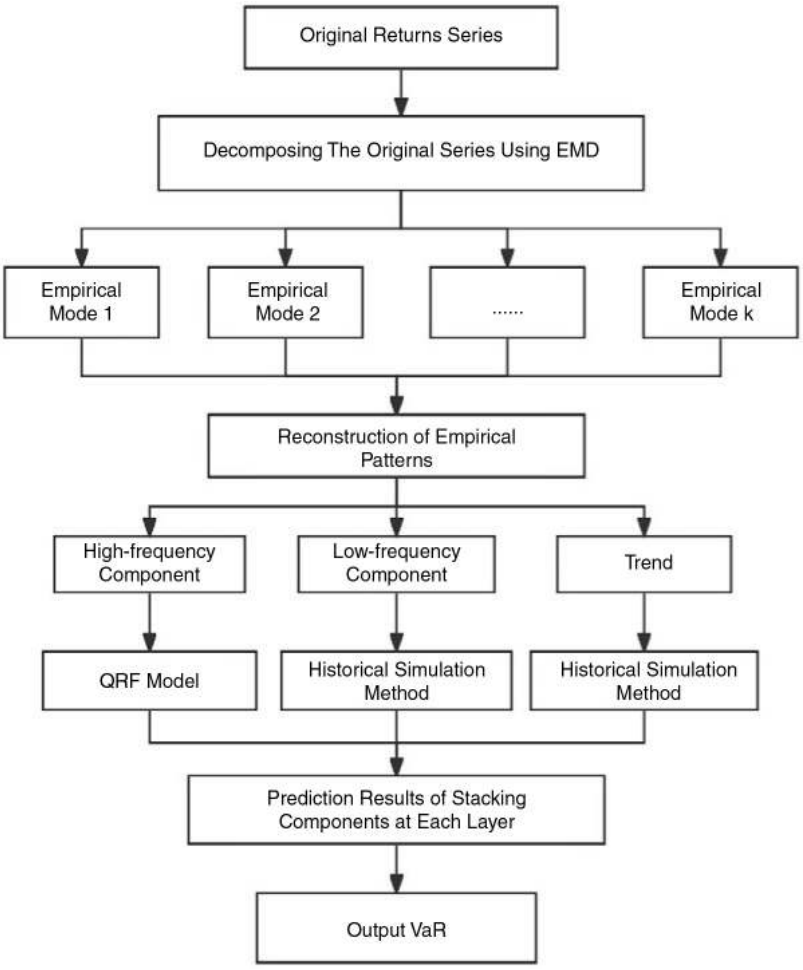


Figure 11.1. Flowchart for the estimation of the VaR of financial assets using the EMD-QRF model.

11.3.1 *Decomposition of the original sequence*

The EMD decomposition method is used to decompose the return series of the CSI 300 Index into a series of IMF components with low correlation and different frequencies, as well as a trend term. The curve of each IMF component is a waveform that is basically symmetric around the zero-mean line defined by its local maxima and minima, representing the oscillating behavior of the original series at different time scales, which can be regarded as different factors affecting the time scale of the returns.

11.3.2 *Reconstruction of empirical mode components*

Based on the correlation, periodicity, and Hurst index of IMF components, the IMFs are divided into high-frequency and low-frequency components, and then the IMF components are reconstructed. The reconstruction method of IMF components is as follows:

$$\text{high-frequency component} = \sum \text{high-frequency term} \tag{11.14}$$

$$\text{low-frequency component} = \sum \text{low-frequency term} \tag{11.15}$$

$$\text{trend component} = \text{Trend} \tag{11.16}$$

While the high-frequency component represents the short-term volatility of the original series, the low-frequency component represents the medium-term volatility of the original series, and the trend item concentrates on reflecting the non-stationarity of the series, representing the long-term trend of the original series.

11.3.3 *Related statistics*

We further explore the economic significance of the reconstructed items by calculating the variance contribution rate, characteristic time scale, and Hurst index of high-frequency components, low-frequency components, and trend components. The following indicators are introduced:

(1) **Variance contribution rate:**

$$\text{STDR} = \frac{\text{Var}(\text{IMF}_i)}{\sum_{i=1}^{k} \text{Var}(\text{IMF}_i)} \tag{11.17}$$

The variance contribution rate indicates the impact of the volatility of a single IMF component on the volatility of the entire series.

(2) **Average period:**

$$C = \frac{N}{N_{\text{Min}}} \tag{11.18}$$

N represents the number of observed values in the original series, and N_{Min} represents the number of extreme values. The average period reflects the time scale of IMF component fluctuations.

(3) **Hurst Index:** The Hurst index is used to judge whether a time series follows a random walk or a biased random walk process. The method for calculating the Hurst index using the R/S method is as follows:

For a time series, divide it into α equal-length sub-intervals of length n. Let M_i denote the mean of the ith interval, then the cumulative deviation of the tth element in the ith interval is given by

$$X_{t,i} = \sum_{u=1}^{t}(x_{u,i} - M_i), \quad t = 1, \ \ldots \ , n \tag{11.19}$$

The range of the *i*th interval is

$$R_i = \max(X_{t,i}) - \min(X_{t,i}) \tag{11.20}$$

Denote the sample standard deviation of the ith interval as S_i, and then the rescaled range of the *i*th sample is given by

$$\left(\frac{R}{S}\right)_n = \frac{1}{A}\sum_{i=1}^{a}\frac{R_i}{S_i} \tag{11.21}$$

If the rescaled range $(R/S)_n$ follows a power law distribution with sub-interval length of n, then the power exponent is the Hurst index, denoted as H:

$$\ln\left(\frac{R}{S}\right)_n = H^{*}\ln n + C, \quad C \text{ is a constant} \tag{11.22}$$

11.4 Empirical Study

11.4.1 *Data source and processing*

The data selected for the empirical study in this chapter was the CSI 300 Index, spanning from January 4, 2016, to February 26, 2021, with a total of 1,252 samples. To assess the stability of the models with historical data, all historical data are divided into two parts: in-sample data and out-of-sample data. The in-sample data consist of the first 974 data points before January 2, 2020, while the out-of-sample data consist of the remaining 278 data points from January 2, 2020, onwards. Logarithmic transformation is applied to the CSI 300 Index return data, and the formula for the logarithmic return at time t is given by

$$r = (\ln P_t - \ln P_{t-1}) * 100\% \tag{11.23}$$

11.4.2 *Logarithmic return decomposition*

The EMD method is utilized to decompose the logarithmic return sequence, resulting in nine IMF components and one trend component, as illustrated in Figure 11.2. The IMF components exhibit decreasing frequencies, and each IMF component is stationary.

To gain a deeper understanding of the statistical characteristics of each IMF component, boxplots for each component (Figure 11.3) and a correlation heatmap (Figure 11.4) were plotted. From Figure 11.3, it can be observed that the median and mean of each IMF component are around zero, and except for IMF5, IMF6, and IMF7 which exhibit asymmetric outliers. The maximum and minimum data points of the other components are symmetrically distributed around zero. Meanwhile, IMF1 to IMF4 exhibit many "+" symbols outside the upper and lower outlier boundaries, indicating that IMF1 to IMF4 are white noise sequences with significant fluctuations. From Figure 11.4, it can be seen that the correlation between the low-frequency component IMF9 and the residual component is the highest at 55%, while the correlation coefficients between other components are relatively low, most below 4%, indicating a lack of correlation. From the above analysis, it can be concluded that EMD can decompose IMFs with different frequencies of high-noise time series from the original series, avoiding noise interference and extracting more valuable information from the original data for research purposes.

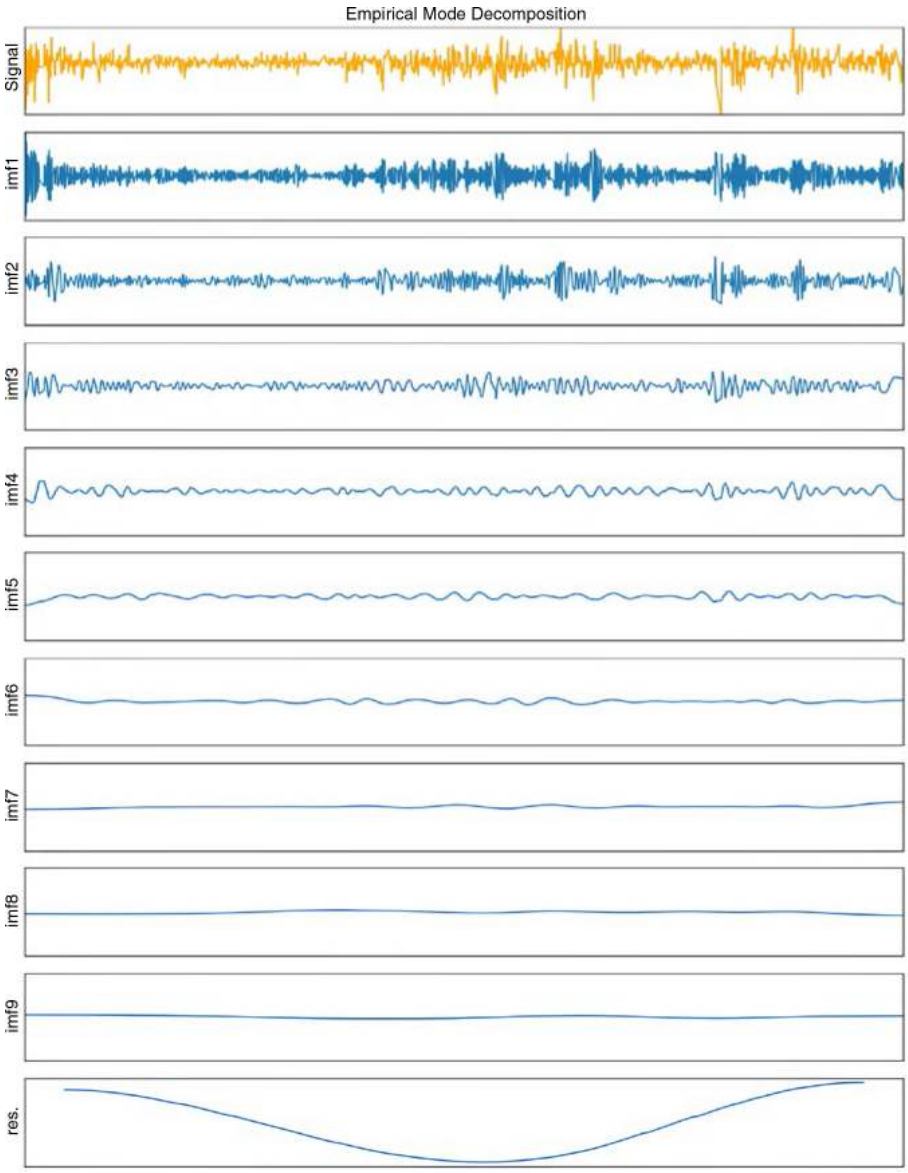


Figure 11.2. EMD decomposition results.

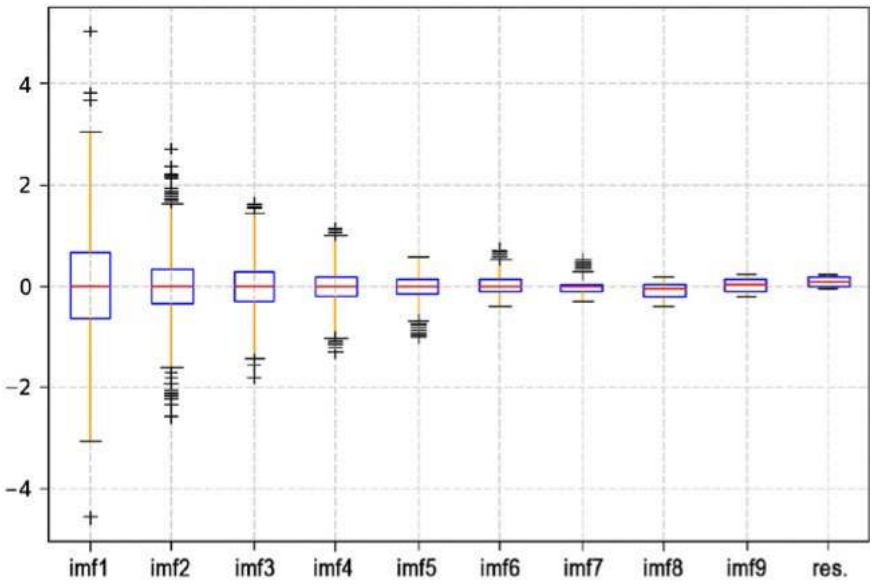


Figure 11.3. Boxplots of each mode component.

The variance contribution rate, average period, and Hurst exponent of each component are shown in Table 11.1. The average periods of IMF1 to IMF4 are stable at below 30 days, with a variance contribution rate of 90.8% and a Hurst exponent around 0.5. IMF5 to IMF8 have average periods which are stable between 30 and 400 days, with a variance contribution rate of 7.4% and a Hurst exponent below 1. IMF9 and the residual component have average periods above 400 days, with a variance contribution rate of 1.8%, and a Hurst exponent exceeding 1.



Figure 11.4. Heatmap of correlation coefficients for each component.

Table 11.1. Statistical characteristics of each component.

Components	STDR	C	H
IMF1	50.5%	2.865	0.467
IMF2	21.2%	5.743	0.385
IMF3	12.4%	11.179	0.412
IMF4	6.7%	22.357	0.488
IMF5	3.2%	39.125	0.588
IMF6	1.9%	78.250	0.691
IMF7	1.1%	165.500	0.826
IMF8	1.2%	313.000	0.972
IMF9	1.1%	417.333	1.005
Res.	0.7%	1252.000	1.006

Based on statistical characteristics such as correlation, average period and Hurst exponent of each mode component, the components can be reconstructed into high-frequency components (Random), low-frequency components (Detail), and trend components (Trend):

$$\text{Random} = \text{IMF1} + \text{IMF2} + \text{IMF3} + \text{IMF4} \tag{11.24}$$

$$\text{Detail} = \text{IMF5} + \text{IMF6} + \text{IMF7} + \text{IMF8} \tag{11.25}$$

$$\text{Trend} = \text{IMF9} + \text{res} \tag{11.26}$$

The trends of the reconstructed components are shown in Figure 11.5. From the figure, it can be observed that the high-frequency component exhibits high-frequency fluctuations, resembling a white noise series with random-walk-like behavior. The low-frequency component represents medium-term cyclical fluctuations in the market, while the trend component appears as a smooth curve. The correlation statistics of each reconstructed component are presented in Table 11.2. From Table 11.2, it can be seen that the variance contribution rate of the high-frequency component is 90.5%, with an average period of three days and a Hurst exponent of 0.364. The variance contribution rate of the low-frequency component is only 6.5%, with an average period of three days and a Hurst exponent of 0.786. From the perspective of the Hurst exponent, the high-frequency component exhibits higher randomness, while the low-frequency component and the trend component demonstrate stronger long-term memory.

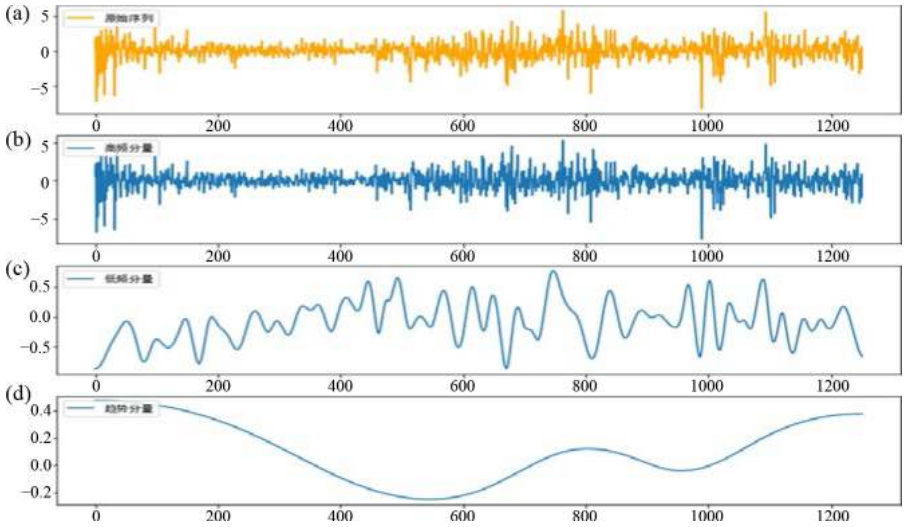


Figure 11.5. EMD decomposition component reconstruction graph. (a) Original series, (b) high-frequency component, (c) low-frequency component, and (d) trend component.

Table 11.2. Reconstructed component statistical characteristics.

Reconstructed components	STDR (%)	*C*	*H*
Random	90.5	3	0.364
Detail	6.5	3	0.786
Trend	3.0	42	0.998

11.4.3 *Model construction*

We employ the QR technique to measure the VaR of the CSI 300 Index. Average returns lagged by five periods are used as explanatory variables. We then consider the impact of volatility on returns and add a new explanatory variable, volatility to the above model. The estimate of volatility is based on realized variance:

$$R_t = \alpha_t + \beta_0 \mathrm{Vol}_t + \beta_1 R_{t-1} + \beta_2 R_{t-2} + \beta_3 R_{t-3} + \beta_4 R_{t-4} + \beta_5 R_{t-5} + \mathrm{e}_t \tag{11.27}$$

Let x denote the average return lagged by five periods, and the estimated value of volatility is

$$\mathrm{Vol} = \frac{(R_{t-1} - \mu_t)^2 + (R_{t-2} - \mu_t)^2 + (R_{t-3} - \mu_t)^2 + (R_{t-4} - \mu_t)^2 + (R_{t-5} - \mu_t)^2}{5} \tag{11.28}$$

QRF is used to calculate VaR. To mitigate the impact of feature selection on estimation results, the same explanatory variables as in the QR model are selected as

input features to estimate the VaR risk measure for CSI 300.

The EMD-QRF model is applied to estimate the VaR of CSI 300. First, the EMD algorithm is used to process the original series, extracting IMF components of different frequencies. Subsequently, for components with similar frequencies, a new component is reconstructed. Then, VaR values for high-frequency components are predicted using QRF, while VaRs for trend and low-frequency components, VaR values are obtained via historical simulation. Finally, the VaR estimates of the three components are aggregated to obtain the VaR value of the original series. The estimation method for the EMD-QR model is similar, except that high-frequency components are estimated using QR instead of QRF.

11.4.4 *Posterior comparison analysis results*

The Kupiec likelihood ratio test method is utilized to quantify the discrepancy between the forecasted value and the actual value of VaR, comparing the four models to determine which model can more accurately predict the VaR of CSI 300 returns. The sample sizes are 974 for in-sample data and 278 for out-of-sample data. Firstly, the obtained VaR series are compared with the actual loss series to obtain different models' failure rates and LR statistics at various confidence levels, as shown in Table 11.3. In the in-sample data, among the four models, the EMD-QRF model exhibits the lowest failure rate, and the LR statistics for EMD-QR, QRF, and EMD-QRF are all significant at 90% and 95% confidence levels. For out-of-sample data, the performance of the EMD-QRF model in terms of failure rate is better than the other three models. Additionally, both QRF and EMD-QRF exhibit significant LR statistics at 90% and 95% confidence levels. This suggests that the QRF is relatively more accurate in characterizing risk compared to other models, and the EMD-QRF demonstrates better estimation performance than the traditional QRF.

In summary, the QRF has an advantage over traditional QR models in VaR estimation. The EMD-QRF model, which combines the EMD algorithm with the QRF, has better generalization capabilities compared to the QRF model.

Table 11.3. Kupiec test results (failure rate/LR).

Sample type	Model	90%		95%		99%	
In-sample	QR	10.4%	0.200	5.1%	0.008	1.1%	0.174
	EMD-QR	14.7%	20.810*	8.2%	17.242*	1.9%	5.764
	QRF	1.4%	119.091*	1.1%	43.706*	1.2%	0.522
	EMD-QRF	1.2%	127.493*	1.1%	43.706*	0.8%	0.313
Out-of-sample	QR	12.5%	1.764	7.4%	2.786	2.2%	2.973
	EMD-QR	11.0%	0.311	7.7%	3.661	1.8%	1.547
	QRF	2.2%	24.752*	1.8%	15.103*	1.5%	−18.193

EMD-QRF	1.8%	29.417*	1.8%	7.477*	1.5%	0.531

Note: $^{*}p = .05$.

11.5 Conclusion

Preventing and resolving financial risks are crucial for achieving high-quality economic development. Effective identification of financial risks is essential for risk prevention, early warning, and regulation. VaR is an important indicator for measuring financial risk. In order to achieve an accurate measurement of VaR, this chapter combines EMD with QRF.

First, the historical return series of the CSI 300 Index is decomposed into multiple IMFs with low correlation and different frequencies, along with a residual component, using the EMD method. Second, the IMFs are reconstructed into high-frequency components, low-frequency components, and trend components based on their correlations, periodicities, and Hurst indices. The high-frequency components exhibit characteristics of high-frequency fluctuations, resembling white noise series with high randomness observed from their Hurst indices. The low-frequency components explain medium-term cyclical fluctuations in the market, while the trend component represents a smooth curve. From the perspective of periodicity and the Hurst index, the low-frequency components and the trend components exhibit strong long-term memory.

Finally, four models, namely QR, EMD-QR, QRF, and EMD-QRF, are applied to measure the VaR of the historical return series of the CSI 300 Index. The Kupiec test results show that the QRF characterizes risk more accurately than the QR model. Moreover, the EMD-QRF model demonstrates better generalization capabilities compared to the traditional QRF, validating the effectiveness and accuracy of this model.

This research demonstrates the effectiveness of VaR measurement methods. By combining EMD with QRF to predict VaR, financial institutions and investors can timely and accurately assess the risks they face, make informed decisions, and take necessary precautions. This helps defend against systemic financial risks, and safeguard the stability of the financial system.

Chapter 12

Summary and Outlook

12.1 Summary

This monograph examines Internet financial platforms by focusing on existing fraud risks in operations, including the identification and prevention of individual fraud and group fraud.

Chapters 1–3 of this monograph introduce the concept of Internet finance and describe the current situation and problems facing the development of Internet finance. Chapters 4–6 discuss several risk assessment models used in Internet finance, including machine learning, deep learning algorithms, and graphical models with a focus on two topics: credit scoring and fraud identification. Then, Chapters 7–9 describe in detail the empirical research we conducted on Internet financial risk control using Random Forest, Text Mining, and Semi-Supervised Adaptive Modeling, respectively. The main goal of our research is to build anti-fraud models to identify individual fraud and group fraud events in Internet finance, and then build a fraud identification and risk control system applied to the credit platform to create a better investment environment for Internet financial investors, and to provide more accurate and comprehensive anti-fraud protection for Internet financial platforms.

The model builds upon existing research and proposes a novel intelligent computational model based on semi-supervised learning. The model is designed to identify fraud and address the the imbalance of labeled data relative to unlabeled data on the Internet finance platforms. This way, the monograph develops a dynamic machine learning model that exhibits improved responsiveness to evolving conditions, thus improving the accuracy of fraud identification.

The monograph constructs an Internet financial dictionary using platform data to identify themes of commentary and text sentiment. It further establishes a public opinion monitoring platform and early warning systems which could identify problematic platforms in a timely and accurate manner. The system could also realize round-the-clock monitoring of operating platforms.

The monograph selects one Internet financial platform as a node and builds a network of platforms around it. Based on network analysis, complex network theory, and text mining of public opinion data, we studied the impact of information dissemination in investment and lending on Internet financial platforms, as well as the correlation between information cross-propagation and platform risk. To do so, we constructed correlation and econometric models and analyzed the mechanisms of risk dissemination among correlated platforms. This study aims to identify and mitigate the risks associated with existing platforms. The study proposes actionable risk mitigation strategies to inform governmental decision-making concerning the prevention and mitigation of Internet financial risks.

Additionally, this monograph presents several methods to evaluate risk associated with investments in Internet finance in Chapters 10 and 11 as a supplement, which aims to provide investors with additional tools to make rational investments.

12.2 Outlook

While many Internet finance platforms in China failed following a wave of collapses in 2018, after years of rapid growth in the mid-2010s, it is important to recognize the positive contributions of the industry—namely, expanding access to financing for underserved individuals and small- and medium-sized enterprises, and offering new investment channels for retail investors. Rather than abandoning the sector due to its elevated risks relative to traditional finance, a more constructive approach is to strengthen regulatory frameworks that support both financial stability and sustainable innovation. This monograph has examined risk control and mitigation in China's Internet finance industry through both the short-term market correction and long-term strategies aimed at fostering healthy and resilient industry development.

12.2.1 *Short-term market correction and risk mitigation*

Following a period of rapid expansion and risk accumulation, China's P2P lending industry underwent a major reshuffling and market correction. Most problematic platforms were forced out of the market through a wave of regulatory interventions, including investigations into fraud, noncompliance, and systemic risk. As these issues attracted public attention, the remaining platforms faced mounting challenges—many struggled with daily operations or experienced a "run on the platform," akin to a traditional bank run. In the P2P context, this refers to investors being concerned with a platform's ability to meet repayment obligations or continue operating, thus causing mass withdrawals that triggered liquidity crises. These crises often resulted in the loss of principal and, in some cases, bankruptcy. The few leading platforms that survived this shakeout—largely due to stronger risk management capabilities—have since transitioned into consumer finance or broader Fintech companies.

For companies that exited the industry due to fraud, noncompliance, or bankruptcy during the market correction, asset liquidation should proceed under regulatory supervision or in accordance with court rulings to ensure orderly repayment of outstanding debts. For those that have successfully transformed their business models, regulators should ensure they are properly classified and licensed according to their risk management and fraud detection capabilities. Appropriate regulatory categorization and oversight are essential to maintaining financial stability and fostering responsible innovation in the evolving Fintech landscape.

12.2.2 *Balancing Risk and Innovation: A Path Forward for Internet Finance Regulation in China*

To promote long-term risk mitigation while supporting innovation, a balanced regulatory framework for Internet finance must be established. First, legislation governing Internet finance should be improved to provide a solid legal foundation for financial innovation and the development of financial technologies. Clear and enforceable legal standards will ensure that innovation can proceed within a well-defined regulatory environment.

Second, the development of a comprehensive credit system should be prioritized. This includes formulating uniform standards and building a

centralized, timely, and complete credit data infrastructure to address the current issue of fragmented “data silos.” Such a system would improve transparency and enhance risk assessment across the industry.

Third, financial innovation must be encouraged while maintaining an appropriate balance between risk prevention and growth. The adoption of a regulatory sandbox model offers a practical solution. As a controlled environment for testing new financial products and services, the sandbox allows for adjustable regulatory parameters—either tightened or relaxed depending on the specific case—thereby promoting innovation without compromising financial stability. This approach can help achieve a win-win outcome where innovation and effective risk control reinforce each other.

References

[1] Yongzu B, Wang Y. Financial public opinion risks in the internet environment. Review of Economic Research, 2017, 36: 25.

[2] Herrero-Lopez, S. Social interactions in P2P lending. In Proceedings of the 3rd Workshop on Social Network Mining and Analysis. Association for Computing Machinery, New York, NY, USA, 2009, Article 3: 1–8.

[3] Emekter R, Tu Y, Jirasakuldech B, Lu M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics. 2015, 47(1), 54–70.

[4] Malekipirbazari M, Aksakalli V. Risk assessment in social lending via random forests. Expert Systems with Applications, 2015, 42: 4621–4631.

[5] Jin Y, Zhu Y. A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending. Proceedings of 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, India, 2015, pp. 609–613, doi: 10.1109/CSNT.2015.25.

[6] Guo Y, Zhou W, Luo C, Liu C, Xiong H. Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 2016, 249: 417–426.

[7] Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. Description-text related soft information in peer-to-peer lending–Evidence from two leading European platforms. Journal of Banking & Finance, 2016, 64: 169–187.

[8] Zhang J, & Liu P. Rational herding in microloan markets. Management Science, 2012, 58(5): 892–912.

[9] Wang H & Liao L. Chinese P2P Platform's Credit Authentication Mechanism Research—Evidence from Renrendai. China Industrial Economics, 2014, 4: 136–147. doi:10.19581/j.cnki.ciejournal.2014.04.011.

[10] Xiao M, Ou Y & Li Y. On the Influence Factors of Credit Risk of Online P2P Lending in China: Based on an Empirical Analysis by the Ranking Selection Model. The Theory and Practice of Finance and Economics, 2015, 36(01):2–6. DOI:10.16339/j.cnki.hdxbcjb.2015.01.001.

[11] Xia Y, Liu C, Xu J. Application of clustering support vector machine in P2P lending default prediction. Proceedings of the 10th (2015) China Academy of Management Annual Conference, 2015: 507–514.

[12] Luo Q, Ding G, Fu X, Cai S & Chen X. Detecting anomaly loans on P2P lending platform: Based on hierarchical classification method. Journal of Industrial Engineering and Engineering Management, 2017, 31(03): 201–209. DOI:10.13587/j.cnki.jieem.2017.03.025.

[13] Freedman S, Jin Z G. The information value of online social networks: Lessons from peer-to-peer lending. International Journal of Industrial Organization, 2017, 51: 185–222. https://doi.org/10.1016/j.ijindorg.2016. 09.002.

[14] Galak, J., Small, D., & Stephen, A. T. Microfinance decision making: A field study of prosocial lending. Journal of Marketing Research, 2011, 48: S130–S137.

https://doi.org/10.1509/jmkr.48.spl.s130

[15] Lin, M., Prabhala, N. R., & Viswanathan, S. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. Management Science, 2013, 59(1): 17–35.

[16] Wu, Y., & Wang, H. Group fraud detection in auction websites based on PageRank and social network analysis. Knowledge Economy, 2013, 15: 48–53.

[17] Wang Y. *Anti-Fraud Analysis and Modeling based on Social and Consumption Data* (Master's Thesis, Lanzhou University), 2017.

[18] Bachmann A, Becker A, Bürckner D, Hilker M, Kock F, Lehmann M, Tiburtius P, Funk B. Online Peer-to-Peer lending: A literature review. Journal of Internet Banking and Commerce, 2011, 16(2): 1–18.

[19] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: The next frontier for innovation, competition, and productivity. Research Technology Management, 2011: 101–102.

[20] Jiang M. Using correlation in big data to improve the accuracy of credit underwriting projections. Shanghai Finance, 2015, 415(2): 100–106.

[21] Lin Y. Risk model of bank bad credit based on big data analytics. Engineering Economy, 2015, 25(6): 112–118.

[22] Chen X. A study of the private borrowing costs — Empirical analysis based on P2P network borrowing and lending. Journal of Finance and Economics, 2014, 29(1): 37–48.

[23] Hunt EB, Marin J, Stone P. Experiments in induction. American Journal of Psychology, 1966, 80(4): 17–19.

[24] Quinlan JR. Induction of decision trees. Machine Learning, 1986, 1: 81–106.

[25] Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1993.

[26] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH. Top 10 algorithms in data mining. Knowledge and Information Systems, 2008, 14: 1–37.

[27] Zhou Z, Yang Q. *Machine Learning and Its Application.* Tsinghua University Press, Beijing, China. 2011, 121–145: 171–196.

[28] Ding S, Qi B, Tan H. An overview on theory and algorithm of support vector machines. Journal of University of Electronic Science and Technology of China, 2011, 40(1): 2–10.

[29] Ding L, Luo P. Research on default risk early-warning in P2P lending based on stacking ensemble strategy. Review of Investment Studies 2017, 36(4): 41–54.

[30] Wolpert DH. Stacked generalization. Neural Networks, 1992, 5: 241–259.

[31] Wiginton JC. A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 1980, 15(3): 757–770.

[32] Nath R, Jackson WM, Jones TW. A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis. Journal of Statistical Computation and Simulation, 1992, 41(1–2): 73–93.

[33] Davis RH, Edelman DB, Gammerman AJ. Machine-learning algorithms for credit-card applications. IMA Journal of Management Mathematics, 1992, 4(1): 43–51.

[34] Lahsasna A, Ainon RN, Teh YW. Credit scoring models using soft computing methods: A survey. The International Arab Journal of Information Technology, 2010, 7(2): 115–123.

[35] Bhattacharyya S, Jha S, Tharakunnel K, Westland JC — *Decision Support Systems*, 2011, 50: 602–613. doi: 10.1016/j.dss.2010. 08.008.

Finlay S. Multiple classifier architectures and their application to credit risk assessment.
[36] European Journal of Operational Research, 2011, 210(2): 368–378.

[37] Yamanishi K, Takeuchi J, Williams G, *et al.* Data mining and knowledge discovery. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms, 2004, 8(3): 275–300.

[38] Yan H, Hu M, Cai R. System design of avoiding credit card fraud. Microcomputer Information, 2006, 22(12): 63–66.

[39] Yang X. *Research on Credit Card Fraud Detection based on Support Vector Machine* (Master's Thesis, Sichuan Normal University), 2008.

[40] Guo T, Li G, Yuan D. Abnormity detection of credit card based on confidence and neural network. Computer Engineering, 2008, 34(15): 205–207.

[41] Yang Y. *Research on Fraud Detection of Credit Card Based on Rare-Class Classification* (Master's Thesis, Beijing Technology and Business University), 2013.

[42] Chen M. Research on credit card fraud detection based on fuzzy local Outlier Factor (LOF). Financial Theory & Practice, 2016, 10: 54–57.

[43] Su Y. *Credit Card Fraud Detection Based on Fuzzy 2-norm Quadratic Surface Support Vector Machine* (Master's Thesis, Dongbei University of Finance and Economics), 2017.

[44] Anis M. *An Imbalanced Approach Towards Credit Card Fraud Detection Using Proximity Based Resampling and Classifier Ranking.* University of Electronic Science and Technology of China, 2018

[45] Collier BC, Hampshire R. Sending mixed signals: Multilevel reputation effects in peer-to-peer lending markets. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work. ACM, 2010: 197–206.

[46] Agarwal S, Hauswald R. Distance and private information in lending. The review of financial studies, 2010, 23(7): 2757–2788.

[47] Meyer T, Heng S, Kaiser S, Walter N. Online P2P lending nibbles at banks' loan business. Deutsche Bank Research, 2007, 2(1): 39–65.

[48] Klafft M. Online peer-to-peer: A Lenders' perspective. Proceedings of the International Conference on E-learning, E-Business, Enterprise Information Systems, and E-Government. EEE, 2008, 371–375.

[49] Chaffee EC, Rapp GC. Regulating online peer-to-peer lending in the aftermath of Dodd-Frank: In search of an evolving regulatory regime for an evolving industry. Washington and Lee Law Review, 2012, (69): 485.

[50] Michels J. Do Unverifiable disclosures matter? Evidence from peer-to-peer lending. The Accounting Review, 2012, 87(4): 36–41.

[51] Xu JJ, Lu Y, Chau M. *P2P Lending Fraud Detection: A Big Data Approach/Intelligence and Security Informatics.* Springer International Publishing, 2015, 9047: 71–81.

[52] Kang S, Cao R. Application study of online big data technology in financing field. Financial Theory and Practice, 2014, 1: 108–110.

[53] Ripley B D. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[54] West D. Neural network credit scoring models. Computers and Operations Research, 2000, 27(11–12): 1131–1152.

[55] Dong S, Wang Q. Research on credit rating of the management headquarters of corporations —based on BP neural network classification algorithm. Journal of ABC University, 2016, 3: 22–25.

[56] Qiao P, Zhu J. The era of artificial intelligence: Reconstruction of China's future accounting credit system. Modern Industrial Economy and Informationization, 2017, 7(13): 50–52.

[57] Wu B, Ye J, Dong M. Research on consumer credit risk assessment model in P2P online lending — based on hybrid fruit fly neural network method. Friends of Accounting, 2017, 21: 32–35.

[58] Li C, Dong Z, Cao X. Credit risk assessment of P2P online lending market based on BP neural network. Modernization of Management, 2015, 35(4): 94–96.

[59] Sirignano JA, Sadhwani A, Giesecke K. Deep learning for mortgage risk. Big Data & Innovative Financial Technologies Research Paper Series, 2016.

[60] Wang C, Han D. Prediction of credit risk in internet financial industry based on convolutional neural network. Microcomputer & Its Applications, 2017, 36(24): 44–46.

[61] Wang Y. *Research on the Individual Credit Risk Control Based on Big Data for Online P2P Lending Platform* (Master's Thesis, Southeast University), 2018.

[62] West D, Dellana S, Qian J. Neural network ensemble strategies for financial decision applications. Computers & Operations Research, 2005, 32(10): 2543–2559.

[63] Kirkos E, Spathis C, Manolopoulos Y. *Data Mining Techniques for the Detection of Fraudulent Financial Statements.* Expert Systems with Applications, 2007, 32(4), 995–1003.

[64] Ala'raj M, Abbod M. A systematic credit scoring model based on heterogeneous classifier ensembles. 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA). IEEE, 2015: 1–7.

[65] Wu B. *The Research of Credit Card Anti-Fraud System Base on Neural Network* (Master's Thesis, Wuhan University of Technology), 2010.

[66] Liu Y. *Study on the SVDD Algorithm and Its Application in Credit Card Fraud Detection* (Master's Thesis, Jiangsu University), 2010.

[67] Ding W. *A Study of Anti-fraud System of Credit Card Transaction Based on Deep Learning Technology* (Master's Thesis, Shanghai Jiao Tong University), 2015.

[68] Xu C. *Research and Implementation of Credit Card Overdue Risk Prediction Based on Deep Learning* (Master's Thesis, Soochow University), 2016.

[69] Fu K. *Credit Card Fraud Detection Using CNN* (Master's Thesis, Shanghai Jiao Tong University), 2017.

[70] Wen L. *Research on Credit Default Identification Method Based on Deep Learning* (Master's Thesis, Beijing University of Technology), 2019.

[71] Chen G. *The KNN-Smote-LSTM Based Credit Card Fraud Risk Detection Network Mode* (Master's Thesis, Zhejiang GongShang University), 2018.

[72] Chen R. *Research on the Detection Model of Credit Card Transaction Fraud Based on GAN-XGBoost* (Master's Thesis, Hangzhou Normal University), 2019.

[73] Zafarani R, Abbasi MA, Liu H. *Social Media Mining: Information Diffusion in Social Media.* Cambridge University Press, 2014.

[74] Ni W. How to use relationship mapping to expose gang fraud in financial scams? Suning Wealth Insights, 2018. https://www.weiyangx.com/306008.html.

[75] Macskassy S, Provost F. *A Simple Relational Classifier.* Technical Report, New York University Stern School of Business, 2003.

[76] Neville J, Jensen D. Iterative classification in relational data. In Proceedings of AAAI-2000 Workshop on Learning Statistical Models from Relational Data, 2000, 13–20.

[77] Sen P, Namata G, Bilgic M, Getoor L, Eliassi-Rad T. Collective classification in network data articles. AI Magazine, 2008, 29(3):93–106.

[78] Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821–7826.

[79] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007, 76(3): 036106.

[80] Xie J, Boleslaw KS. Label Rank: A stabilized label propagation algorithm for community detection in networks. In *2013 IEEE 2nd Network Science Workshop (NSW)*. IEEE, 2013: 138–143.

[81] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014:701–710.

[82] Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 855–864.

[83] Ribeiro, L. F. R., Saverese, P. H. P., & Figueiredo, D. R. struc2vec: Learning node representations from structural identity. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 385–394.

[84] Kipf T N and Welling M. *Semi-supervised Classification with Graph Convolutional Networks.* arXiv preprint arXiv:1609.02907, 2016.

[85] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems, 2017: 1024–1034.

[86] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, and Bengio Y. *Graph Attention Networks*. arXiv preprint arXiv:1710.10903, 2017.

[87] Van Vlasselaer V, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B. Gotcha! Network-based fraud detection for social security fraud. Management Science, 2016, 63(9): 3090–3110.

[88] Tseng VS, Ying J, Huang C, Kao Y, & Chen K. FrauDetector: A graph-mining-based framework for fraudulent phone call detection. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 2157–2166.

[89] Lismont J, Ram S, Vanthienen J, Lemahieu W, Baesens B. Predicting interpurchase time in a retail environment using customer-product networks: An empirical study and evaluation. Expert Systems with Applications, 2018, 104: 22–32.

[90] Zhan Q, Yin H. A loan application fraud detection method based on knowledge graph and neural network. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence. ACM, 2018: 111–115.

[91] Van Vlasselaer V, Bravo C, Caelen O, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B. Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 2015, 75: 38–48.

[92] Lebichot B, Braun F, Caelen O, Saerens M. A Graph-based, semi-supervised, credit card fraud detection system. In International Workshop on Complex Networks and their Applications. Springer, 2016: 721–733.

[93] Zhao Y, Xie Y, Yu F, Ke Q, Yu Y, Chen Y, Gillum E. BotGraph: Large scale spamming botnet detection. In NSDI, 2009, 9: 321–334.

[94] Wang J, Wen R, Wu C, Huang Y, Xion J. FdGars: Fraudster detection via graph convolutional networks in online app review system. In Companion Proceedings of the 2019 World Wide Web Conference, ACM, 2019, 310–316.

[95] Ma J, Zhang D, Wang Y, Zhang Y, & Pozdnoukhov A. GraphRAD: a graph-based risky account detection system. In Proceedings of ACM SIGKDD conference. August, 2018, 9: 1–9.

[96] Fan C, Wang L, Xie M. Research on the Risk Identification of P2P Internet Lending in New Economy. Statistical Research. 2017, 2: 35–45.

[97] Ling CX, Huang J, Zhang H. AUC: A better measure than accuracy in comparing learning algorithms. In Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceeding 16 (pp. 329–341). Springer Berlin Heidelberg.

[98] Yoon Y, Li Y, Feng Y. Factors affecting platform default risk in online Peer-To-Peer (P2P) lending business: An empirical study using Chinese Online P2P Platform Data. Electronic Commerce Research, 2019, 19: pp. 131–158.

[99] Liu H, Lin B, Yun W, Ding Y. Comparison of balancing methods in interne finance overdue recognition: Taking PPDai.com as case. Computer Science, 2019, 46(11A): 595–598.

[100] Li W, Ding S, Chen Y. Heterogeneous ensemble for default prediction of Peer-to-Peer lending in China. IEEE Access, 2018, 6: 54396–54406.

[101] Dietterich TG. Machine learning research: Four current directions. Artificial Intelligence Magazine, 1997, 4: 97–136.

[102] Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers. Connection Science, 1996, 8(3–4): 385–404.

[103] Zhou ZH, Li M. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529–1541.

[104] Li M, Zhou ZH. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Transactions on Systems Man and Cybernetics — Part A Systems and Humans, 2007, 37(6): 1088–1098.

[105] Han Q, Zhang Y, Niu W. Microblogging water army identification based on semi-supervised collaborative training algorithm. Computer Science, 2019, 46(11): 202–208.

[106] Dai W, Hu J, Cheng Y, Wang X, Chai T. RVFLN-based online adaptive semi-supervised learning algorithm with application to product quality estimation of industrial processes. Journal of Central South University 2019, 26(12): 3338–3350.

[107] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of International Conference on Machine Learning, 2017: 1126–1135.

[108] Li X, Long S, Zhu J. Survey of few-shot learning based on deep neural network. Application Research of Computers, 2020, 37(8): 2241–2247.

[109] Abbasi A, Albrecht C, Vance A, and Hansen J. MetaFraud: A meta-learning framework for detecting financial fraud. MIS Quarterly, 2012 36(4): 1293–1327.

[110] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 2001, 29(5): 1189–1232.

[111] Vinayak RK, Gilad-Bachrach R. Dart: Dropouts meet multiple additive regression trees. Artificial Intelligence and Statistics, PMLR, 2015, (Vol. 38, pp. 489–497).

[112] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A highly efficient gradient boosting decision tree. In 31s Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, December 4–9, 2017, pp. 1–9.

[113] Engle RF. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom Inflation. Econmetrica, 1982, 50(4): 987–1007.

[114] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 1986, 31(2): 307–327.

[115] Li Y, Xue J. Research on stock index futures risk based on GARCH-VaR model. Chinese Business & Trade, 2013, 23: 79–81

[116] Nelson DB. Conditional heteroskedasticity in asset returns: A new approach. Econometrica: Journal of the Econometric Society, 1991, 59(2): 347–370.

[117] Engle R, Lilien D, Robins R. Estimating time varying risk premia in the term structure: The Arch-M model. Econometrica, 1987, 55(2): 391–407.

[118] Rabemananjara R, Zakoian JM. Threshold ARCH models and asymmetries in volatility. Journal of Applied Econometrics, 1993, 8(1): 31–49.

[119] Yu JX, Wang XH. Volatility prediction and VaR metrics of realized GARCH model based on mixing. Statistical Research, 2018, 35(1): 104–116.

[120] Ma Y, Shao X, Huang B, Tao T. Risk analysis of international oil price fluctuation based on SWARCH model. Statistics and Decision, 2019, 35(14): 170–173.

[121] Koenker R, Bassett G. Regression quantiles. Econometrica, 1978, 46(1): 33–50.

[122] Li B. Risk Analysis of Soybean futures market based on quantile regression. Chinese Business Theory, 2017, 33: 9–11.

[123] Bayer S. Combining value-at-risk forecasts using penalized quantile regressions. Econometrics and Statistics, 2018, 8: 56–77.

[124] Liu YS. *Bayesian Quantile Regression Model and Its Application in Financial Risk Management*. Shan Dong University, 2017.

[125] Meinshausen N, Ridgeway G. Quantile regression forests. Journal of Machine Learning Research, 2006, 7(6): 983–999.

[126] Park S, Kwon J, Kim J, & Oh HS. Prediction of extremal precipitation by quantile regression forests: from SNU Multiscale Team. Extremes, 2018, 21: 463–476.

[127] Sun G, Liang Z, Yu N, Ni X, Wei Z, Zang H, Zhou Y. Short-term wind power probability density prediction based on EWT and quantile regression forest. Power Automation Equipment, 2018, 38(8):165–172.

[128] Chen M Y, Chen J E. Application of quantile regression to estimation of value at risk. Review of Financial Risk Management, 2002, 1(2): 15.

[129] Puneet P, Vikas S, Kewal S. Transformational approach to analytical value-at-risk for near normal distributions. Journal of Risk and Financial Management, 2021, 14(2): 51.

[130] Youssef M, Belkacem L, Mokni K. Value–at–risk estimation of energy commodities: A long–memory GARCH–EVT approach. Energy Economics, 2015, 51: 99–110

[131] Wozniak T. Testing causality between two vectors in multivariate GARCH models. International Journal of Forecasting, 2015, 31(3): 876–894.

[132] Wang X, Cai J, Tang L, He K. VaR measurement for stock portfolio based on BEMD-Copula-GARCH model. Systems Engineering Theory & Practice, 2017, 37(2): 303–310.

[133] Valizadeh, T., Rezakhah, S., & Mohammadi Basatini, F. On time-varying amplitude HGARCH model. International Journal of Finance & Economics, 2021, 26(2): 2538–2547.

[134] Fan J and Gu J. Semiparametric estimation of Value at Risk. Econometrics Journal, 2003 6: 261–290.

[135] Lyu Y, Fu T, Hu Y, Dai D. A study of risk measurements of Chinese gold market based on Bootstraped filtered historical simulation approaches. Chinese Journal of Management Science, 2019, 27(07): 46–55.

[136] Chi G, Zhang Y, Ding S. Loan Portfolio selection model based on power spectral risk measure and Monte Carlo simulation. Chinese Journal of Management Science, 2019, 27(9): 1–14.

[137] Koenker R, Bassett G. Regression quantiles. Econometrica, 1978, 46(1): 33–50.

[138] Wang X, Zhao S. Quantile regression based risk measurement in the Chinese stock markets. Journal of China University of Mining & Technology, 2008, 37(3): 416–421.

[139] Bayer S. Combining value-at-risk forecasts using penalized quantile regressions. Econometrics and Statistics, 2018, 8: 56–77.

[140] Liu Y. Bayesian Quantile Regression and Its Application in Financial Risk Management. (Master's Thesis, Shandong University), 2017.

[141] Xu Q, Xu J, Jiang C, Liu X. Financial risk measure of VaR based on quantile regression neural network. Journal of Hefei University of Technology (Natural Science), 2014, 37(12): 1518–1522.

[142] Xu Q; Zhang J, Jiang C. Evaluating multiperiod VaR via support vector quantile regression. Journal of Systems Engineering, 2014, 29(2): 202–214.

[143] Sun W, Wang F. Study of VaR calculation methods based on least squares support vector machine. Value Engineering, 2017, 36(6): 212–215.

[144] Fang K. *Random Forest Combination Prediction Theory and Its Application in Finance.* Xiamen, China, Xiamen University Press, 2012: 187–205.

[145] Gou X, Wang Q. VaR estimation based on quantile regression forest and risk factors analysis. Journal of University of Science and Technology of China, 2019, 49(8): 635–644.

[146] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non–stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903–995.

[147] Lizhen L. Internet finance: Mechanism, characteristics, supervision, trends. Contemporary Economic Research, 2014, 12: 37.

[148] Zeng X. Surveying unemployment rates is a scientific measure to promote higher quality and more sufficient employment. National Bureau of Statistics, 2018. https://www.stats.gov.cn/xxgk/jd/sjjd2020/201804/t20180420_1764615.html

[149] Kupiec P. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives, 1995, 3(2): 73–84.

Index

A

abstract extraction, 148–149
academic credential, 129
accuracy, 132–134, 139–140, 148, 161–165, 179–180
adaptive learning, 189–192, 199
adaptive model, 10
adaptive semi-supervised algorithm, 193
adverse selection, 25
Alipay, 12, 15, 18–19, 23, 28–29, 36–38
anti-fraud, 8–10, 61, 63, 72
APATE, 99–100, 102
Apriori, 123–125, 134, 140, 146
Area Under the Curve (AUC), 132–134, 136, 139–140, 161–165, 179–180
association rule, 123
attention mechanism, 92–93, 113
Autoregressive Conditional Heteroskedasticity (ARCH), 201
average period, 226, 229–231

B

Bagging, 117, 119–120, 126
balancing, 178–180
banking system, 41
betweenness, 83, 95
big data, 1, 3–5, 8–9, 21–23, 42–45, 49–50, 57
borrower default, 11, 25
bot-user, 103–105
botnet, 103

C

class imbalance, 162, 178
classifier, 72, 184, 186–187, 193, 196, 199
cloud computing, 21–22
cloud technology, 1
cluster, 80, 111
CNN-feature, 192, 196–198
collective classification, 77–78
collective fraud risk, 77
community, 76, 80–83, 110–111

community detection, 53, 81–83, 110
community extraction, 111
community mining, 76–77, 80–81, 113
community search, 53
conditional volatility, 210
connection pattern, 82
convolutional computation, 89
Convolutional Neural Network (CNN), 64
corpus, 151
credit card fraud, 73
credit default identification, 56–58
credit scoring system, 46
CreditX, 96, 102
crowdfunding, 5, 11–12, 15–18, 27, 32–35, 39
    equity-based platform, 33
    philanthropy-focused platform, 33
    property-based platform, 33
    reward-based platform, 33
CSI 300 index, 202, 208

D

DART, 185, 190, 196, 198–199
data mining, 4–5
decision tree, 4, 117–120, 134, 161, 206
decomposition efficiency, 220
DeepWalk, 84–86, 88
dynamic influence process, 78

E

e-commerce, 5, 12, 14–15, 19
edge, 83, 103–104, 107, 113
EMD-QRF, 217, 220, 223–224, 232–234
empirical mode decomposition (EMD), 220
    method, 220
    algorithm, 220
ensemble learning, 183–184, 199
equity crowdfunding, 12, 17–18
expert feature, 102

F

F1 Score, 161–162, 164–166, 179–180
feature combination, 123–125, 134, 136
feature construction, 124–125, 157, 177, 179
feature engineering, 124
feature extraction, 94–96, 99–100, 102–103
feature pruning, 148
feature screening, 124, 134, 136
financial technology (Fintech), 21

financial time series, 201, 203–204, 216
fraud, probability of, 37
fraud detection, 46, 50–51, 53, 55–57
fraud identification, 1–2, 4, 9
fraud prevention, 2, 9
fraud warning system, 182
frequent pattern, 123, 134

G

gain rate, 117–119
GARCH-VaR, 201
GATs algorithm, 92
Gaussian mixture model, 2
Generalized Autoregressive Conditional Heteroskedasticity (GARCH), 201–203, 210–212, 215–216, 218
generative adversarial network, 73
Gini Index, 119
Girvan–Newman algorithm, 54
global feature, 96–98, 100, 102, 108
GOSS, 185, 190, 196, 198–199
GotCha, 96–100, 102
Gradient Boosting Decision Tree (GBDT), 161, 185, 190, 198–199
graph-based semi-supervised learning, 89
graph neural network, 84, 89, 110, 112
graph representation model, 84, 105
graphical model, 10, 75
GraphRAD, 110
GraphSAGE, 84, 91–92
group fraud, 45–46, 50–51, 53

H

herding effect, 3, 4
high-dimensional, 122
high-frequency component, 223, 225–226, 230–232, 234
Hurst exponent, 229–231
Hurst index, 225–227, 234

I

imbalance of the dataset, 126
individual fraud, 50–51
information asymmetry, 31, 38
information gain, 117–119
information security, 43
intelligent customer service, 62–63
Internet finance, 11, 14–17, 20–21, 23–26
Internet financial dictionary, 236
Internet wealth management, 20–21
intrinsic mode function (IMF), 220

investment recommendation algorithm, 142, 146
investment-efficient frontier, 142
isolation forest (IF), 52
iterative classifier, 79

K

K-nearest neighbor (KNN), 106, 148, 161
kernel regression, 49, 122, 141–146
Kupiec likelihood ratio test, *see also* Kupiec test, 214–215, 233

L

label propagation, 113
Label Propagation Algorithm (LPA), 83–84
Laplacian Kernel, 122
LDA, 149, 154–155
Lending Club, 2, 8, 116
LiCaiTong, 36
Lift, 196, 198
LightGBM, 187, 191–192
loan cancellation, 129, 131
local feature, 102, 108
local outlier factor (LOF), 51
logarithmic return, 227
logistic regression (LR), 2, 4, 47, 50, 55–56, 129, 132, 134, 139–140, 215
long short-term memory (LSTM), 64
loopy belief propagation, 79–80
Louvain algorithm, 54–55
low density separation (LDS), 185
low-frequency component, 223, 225–226, 229–232, 234

M

mobile payment, 6, 19–20, 23
modularity, 81
moral hazard, 25

N

Naïve Bayes (NB), 161
naïve classification, 78
neighboring node, 93–94
network security, 37–38
neural network, 64–65, 67–68, 70–72
node, 76–80, 82–84, 86–91, 93–97, 99, 104, 106, 109, 111–113
node classification, 76–77, 89, 94, 112–113
node scoring, 111
Node2Vec, 84, 86–88, 106, 108
non-financial institution, 18, 23
non-linear time series, 220

non-overlapping communities, 82, 110
non-parametric method, 217–218
non-parametric regression, 122
non-performing loan, 25
non-stationary time series, 220

O

O2O credit model, 16
overdue rate, 125–126
overfitting, 132
overlapping communities, 82, 110
oversampling, 73

P

P2P lending platform, 31
P2P online lending, 11, 15–16, 21
Page algorithm, 96
Paipaidai, 11, 15–16, 116, 126, 129, 131, 141, 143–144
parametric method, 217
partition, 83
PayPal, 12
Peer-to-Peer (P2P) lending, 3, 115, 125–126, 138
personal information, 37
personalized PageRank algorithm, 95–96
platform credit risk, 3
platform fraud, 170, 173–176, 181
platform risk, 236
polynomial Kernel, 122
prediction lag, 176
product crowdfunding, 12, 17
Prosper, 2–3
public opinion, 236

Q

QR code payment, 27–29
Qschou, 18
quantile regression forest (QRF), 202–203, 205, 213, 215–216, 222
quantile regression (QR), 201–205, 210–213, 215–216, 218–219, 221

R

radial basis function (RBF) Kernel, 122
random forest, 49–50, 56, 115–117, 120–121, 132–134, 139–140, 142–146, 219, 222
random walk, 226
recall rate, 165–166, 179–180
Recency–Frequency–Monetary Value (RFM), 99
recurrent neural network (RNN), 64, 89
regulation, 19, 24

reinvestment, 143–144, 146
risk control, 61, 63–64, 66–69, 71
robo-advisory, 60, 62

S

sample imbalance, 72
scorecard model, 48
Security First Network Bank, 14
self-sampling, 119
semi-GCN method, 90
semi-parametric method, 218
semi-supervised, 10
semi-supervised clustering, 187
semi-supervised deep learning, 73
semi-supervised dimensionality reduction, 187
semi-supervised learning (SSL), 183–186, 235
semi-supervised regression, 187
sentiment analysis, 151, 153–154, 168, 175
Shanghai Stock Exchange Composite Index (SSEI), 210
simulation, 142–144, 146
Skip-Gram, 84–86, 88
SMOTE, 73
social credit data, 43
social network, 3–5, 7–8
social relationships, 3–4
standard deviation, 122, 141–142
static network, 81–82
statistical method, 3
stock index futures, 201, 202–203, 207–210, 212–214, 216
Struc2Vec, 84, 87–88
supervised learning, 151
support vector machine (SVM), 3–4, 50, 56–57, 148, 165–166

T

Tenpay, 19, 28–29
text classification, 148–149
text data mining, 235–236
text mining, 2, 10, 147–149, 151
TF-IDF, 154
third-party payment, 5, 27
time series data modeling, 64
topic extraction, 148–149, 154, 168, 175
trend component, 224–227, 231, 234

U

UnionPay, 19, 29
unlabeled node, 77–79, 83, 89, 95
user behavior, 4–5

V

Value at Risk, (VaR), 201–205, 210–217
variance contribution rate, 225–226, 229, 231

W

wealth management, mobile, 5
WeChat Pay, 12, 19

X

XGBoost, 47–48

Y

Yu’e Bao, 13, 15, 18, 20–21, 36

Z

Zopa, 11, 15, 115