INSTITUTO
UNIVERSITÁRIO
DE LISBOA

An Empirical Study on Credit Evaluation of SMEs Based on Detailed Loan Data

CHENG Zheng

Doctor of Management

Supervisors:
PhD Vasco Barroso Gonçalves, Assistant Professor,
ISCTE University Institute of Lisbon
PhD Sérgio Moro, Associate Professor with Habilitation,
ISCTE University Institute of Lisbon
PhD LIU Bo, Professor,
University of Electronic Science and Technology of China

May, 2022

iscte

BUSINESS
SCHOOL

Marketing, Operations and General Management Department

An Empirical Study on Credit Evaluation of SMEs Based on Detailed Loan Data

CHENG Zheng

Doctor of Management

Supervisors:
PhD Vasco Barroso Gonçalves, Assistant Professor,
ISCTE University Institute of Lisbon
PhD Sérgio Moro, Associate Professor with Habilitation,
ISCTE University Institute of Lisbon
PhD LIU Bo, Professor,
University of Electronic Science and Technology of China

May, 2022

iscte

BUSINESS
SCHOOL

Marketing, Operations and General Management Department

An Empirical Study on Credit Evaluation of SMEs Based on Detailed Loan Data

CHENG Zheng

Doctor of Management

Jury:
PhD Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE - Instituto Universitário de Lisboa
PhD Carlos Manuel Jorge da Costa, Associate Professor,
ISEG - Universidade de Lisboa
PhD Li Qiang, Ful Professor,
University of Electronic Science and Technology of China
PhD Nelson José dos Santos António, Emeritus Professor,
ISCTE - Instituto Universitário de Lisboa
PhD Vasco Barroso Gonçalves, Assistant Professor,
ISCTE - Instituto Universitário de Lisboa

May, 2022

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

An Empirical Study on Credit Evaluation of SMEs Based on Detailed Loan Data

CHENG Zheng

Declaration

I declare that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university and that to the best of my knowledge it does not contain any material previously published or written by another person except where due reference is made in the text.

Signed: CHENG ZHENG Date: 2022.06.30

Name: CHENG Zheng

作者申明

本人郑重申明：除了论文致谢中明确说明并致以谢意的部分外，所呈交的论文不包含任何他人或作者本人已用于获得任何教育机构的学位和证书而使用过的材料。同时尽我所知，除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。

作者签名：CHENG ZHENG 日期：2022.6.30

姓名(拼音)：CHENG Zheng

[This page is deliberately left blank.]

Abstract

Small and micro-sized Enterprises (SMEs) are an important part of Chinese economic system.The establishment of credit evaluating model of SMEs can effectively help financial intermediaries to reveal credit risk of enterprises and reduce the cost of enterprises information acquisition. Besides it can also serve as a guide to investors which also helps companies with good credit.

This thesis conducts an empirical study based on loan data from a Chinese bank of loans granted to SMEs. The study aims to develop a data-driven model that can accurately predict if a given loan has an acceptable risk from the bank's perspective, or not. Furthermore, we test different methods to deal with the problem of unbalanced class and uncredible sample. Lastly, the importance of variables is analyzed. Remaining Unpaid Principal, Floating Interest Rate, Time Until Maturity Date, Real Interest Rate, Amount of Loan all have significant effects on the final result of the prediction.The main contribution of this study is to build a credit evaluation model of small and micro enterprises, which not only helps commercial banks accurately identify the credit risk of small and micro enterprises, but also helps to overcome creditdifficulties of small and micro enterprises.

Keywords: Credit Evaluation for SMEs; Detailed Loan Data; Machine Learning
JELClassification:C38, C52

[This page is deliberately left blank.]

Resumo

As pequenas e microempresas constituem uma parte importante do sistema económico chinês. A definição de um modelo de avaliação de crédito para estas empresas pode ajudar os intermediários financeiros a revelarem o risco de crédito das empresas e a reduzirem o custo de aquisição de informação das empresas. Além disso, pode igualmente servir como guia para os investidores, auxiliando também empresas com bom crédito.

Na presente tese apresenta-se um estudo empírico baseado em dados de um banco chinês relativos a empréstimos concedidos a pequenas e microempresas. O estudo visa desenvolver um modelo empírico que possa prever com precisão se um determinado empréstimo tem um risco aceitável do ponto de vista do banco, ou não. Além disso, são efetuados testes com diferentes métodos que permitem lidar com os problemas de classes de dados não balanceadas e de amostras que não refletem o problema real a modelar. Finalmente, é analisada a importância relativa das variáveis. O montante da dívida por pagar, a taxa de juro variável, o prazo até a data de vencimento, a taxa de juro real, o montante do empréstimo, todas têm efeitos significativos no resultado final da previsão. O principal contributo deste estudo é, assim, a construção de um modelo de avaliação de crédito que permite apoiar os bancos comerciais a identificarem com precisão o risco de crédito das pequenas e micro empresas e ajudar também estas empresas a superarem as suas dificuldades de crédito.

Palavras-chave:Avaliação do Crédito das Pequenas e Micro Empresas; Dados Detalhados de Empréstimos; Aprendizagem Automática

JELClassificação:C38, C52

[This page is deliberately left blank.]

摘要

小微企业（SME）是中国经济体系的重要组成部分。建立小微企业信用评价模型可以有效地帮助金融中介机构揭示企业信用风险，降低企业信息获取成本。此外，它还可以为投资者提供指导，也可以帮助信誉良好的公司更好的融资。

本文基于一家中国银行 2011-2016 年小微企业贷款数据进行了实证研究。旨在通过构建一个数据驱动的模型，从银行角度预测特定贷款的可接受风险水平。此外，本文采用了多种方法进行测试，以解决不平衡数据集和样本可信度低问题。最后，本文分析了相关因素的影响，发现剩余未偿还本金、浮动利率、至到期日时间、实际利率贷款金额都对预测结果有重要影响。本文的主要贡献在于构建了小微企业信用评估模型，不仅有助于商业银行准确识别小微企业信用风险，也有助于克服小微企业“信贷难”问题。

关键词：小微企业信用评价；贷款明细；机器学习

JEL 分类号:C38, C52

[This page is deliberately left blank.]

Acknowledgements

When I was about to finish my doctoral thesis, looking back on my time in UESTC, I could not grow without the encouragement and help of my teachers, colleagues, relatives and friends. Thank you sincerely.

First of all, I need to thank mytwotutora Professor Liu Bo and Associate Professor Vasco B Gonçalves. The two professors are knowledgeable and rigorous. They have carefully read and corrected this study time and time again, and urged me to revise it again and again. The completion of this studyis inseparable from their guidance and supervision. Their teaching is the most valuable wealth I have obtained during my study.

Secondly, I want to thank my family in particular. During my doctoral study, I was faced with the triple pressure of study, work and life. They were my strong backing, gave me care and support, understanding and tolerance, and gave me the courage and motivation to move forward. Finally, I would like to thank my colleagues for their help and support in my work over the past few years.

I would like to express my heartfelt thanks to all the professors, relatives and friends who have helped me with this research!

[This page is deliberately left blank.]

致谢

博士论文即将完成之际，回首在电子科技大学度过的时光，我的每一步成长离不开老师、同事和亲友们的鼓励与帮助，衷心感谢!

首先需要感谢我的导师刘波教授和 Vasco B Gonçalves 副教授。二位老师学识渊博，治学严谨。一次次精心阅读与斧正本文，并督促促我反复修改，本文的完成离不开他们的指引与督导,他们的教导是我求学期间获取的最宝贵财富!

其次要特别感谢我的家人们。在读博期间，面临学业、工作与生活的三重重压，他们是我坚强的后盾，给予我关心和支持、理解与包容，给予我不断前进的勇气和动力。最后也要感谢我的同事们，在过去的数年里一直给予我工作上的帮助与支持。

仅以此论文向所有帮助过我的老师、亲人和朋友们表示最衷心的谢意!

[This page is deliberately left blank.]

Contents

[This page is deliberately left blank.]

Listof Tables

[This page is deliberately left blank.]

List of Figures

[This page is deliberately left blank.]

List of Abbreviations

	Abbreviation
Analytic hierarchy process	AHP
Area Under Curve	AUC
Back Propagation	BP
False Negative	FN
False Positive	FP
False Positive Rate	FPR
Gross Domestic Product	GDP
K-Nearest Neighbor Classification	KNN
L1-norm penalty	L1
L2-norm penalty	L2
Ministry of Industry and Information Technology	MIIT
Precision rate	P
Principal Component Analysis	PCA
Precision Rate - Recall Rate curve	PR curve
People Repulic of China	PRC
Recall rate	R
Receiver Operating Characteristic Curve	ROC curve
Small and Micro-sized Enterprises	SME
Synthetic Minority Oversampling Technique	SMOTE
Support Vector Machine	SVM
True Negative	TN
True Positive	TP

[This page is deliberately left blank.]

Chapter 1: Introduction

1.1 Background, object and significance

1.1.1 Background

Small and micro-sized Enterprises(SMEs) are the largest and most innovative enterprise groups in China, especially in the era of mass entrepreneurship and innovation, and SMEs are playing an increasingly important role in Chinese economic system to stabilize economic growth, promote transformation and upgrading, stimulate entrepreneurial innovation, provide jobs and improve people's wellbeing(Lv, 2015).

The Ministry of Industry and Information Technology (MIIT) of PRC said that small and medium-sized enterprises currently provide more than 50% tax revenue, 60% GDP, 70% patent inventions and 80% urban employment,and accounting for 90% of business entities. With the gradual advance of policies, such as the ‘Made in China 2025 ‘and the ‘Internet + Strategy’, the important role of SMEs in building strong manufacturing and networking country will be further highlighted (Lv, 2015).

In recent years, the government and financial regulatory authorities have been aware of the financing difficulties of small and micro enterprises, but the proportion of bank loans in the capital sources of small and micro enterprises is still small, and the credit difficulties of small and micro enterprises are common. In order to obtain the cash flow to maintain production and operation, many small and micro enterprises can only rely on their own strength to finance. According to the statistics of relevant institutions, more than half of the capital sources of small and micro enterprises depend on internal financing composed of free capital and retained earnings, and even private usury financing through informal channels. Among them, self-owned funds account for about 30% of the total capital sources of small and micro enterprises, and internal retained funds account for about 26%. However, it is difficult for small and micro enterprises to obtain external financing, especially external equity financing and corporate bonds, accounting for less than 1% of their capital sources. The lack of financing channels has seriously restricted the growth and potential development of production and operation of small and micro enterprises.

As the most important link in the financial system, commercial banks have played an important role in solving the financing difficulties of small and micro enterprises. However, it is generally believed that commercial banks have a lot of room to improve in providing credit services for small and micro enterprises. Objectively speaking, in recent years, many commercial banks have realized the necessity of adjusting the credit structure and launched many innovative financial products with small and micro enterprises as the main service objects, such as Industrial Bank's "Easy Loan", "Zhaodai" of China Merchants Bank, and "CCB benefits you" of China Construction Bank. However, in the credit business of small and micro enterprises, the Internet financial enterprises have the largest market share, of which the number of users of online commercial banks is as high as 29 million, far exceeding the 1590700 users of the second China Construction Bank.

At present, the credit business of small and micro enterprises is still a "blue ocean market", with relatively weak horizontal competition and relatively high profit margin. It is expected to become a new profit growth point of commercial banks. However, the reason that restricts the credit business of small and micro enterprises from becoming the emerging businessof commercial banks is the credit evaluation mechanism of small and micro enterprises. First, small and micro enterprises are small in scale and lack core competitiveness, and the market risk of financial intermediaries providing loans to them is too high; Secondly, small and micro businesses have asymmetric information problems, such as small and micro enterprises' financial standardization and internal control management capacity, which leads to the high cost of financial management of loans from small financial enterprises to small and micro businesses(Guo, 2013). Compared with the complete small and micro enterprise service system of foreign commercial banks, there is still much room for the development of small and micro financial business of commercial banks in China. Therefore, how to build a credit evaluation system for small and micro enterprises, effectively control risks, adhere to the credit approval system, and organically combine the effective allocation of credit resources with the maximization of bank income. It is a problem that must be solved in promoting the development of financial business of small and micro enterprises. At the same time, this is also a practical problem that commercial banks must face to fulfill their social responsibility and effectively support the operation of small and micro enterprises(N. N. Meng&Li, 2018).

At present, the downward pressure on the economy is increasing, and the risk preference of banking financial institutions tends to be cautious. The essence of the problem of "financing difficulty" of small and micro enterprises is the information asymmetry between banks and enterprises. Banks lack clear judgment on the quality of financial information of small and micro enterprises, leading to the adverse choice of credit extension. "Not daring to lend, not willing to lend" has become their dominant strategy. By using machine learning methods, we can overcome the shortcomings of traditional credit risk assessment models in identifying risks due to the small sample size and the non normal distribution of errors. By objectively assessing the credit risk exposure of small and micro loan enterprises, we can not only break through the limitations of traditional risk assessment classification models, meet the requirements of bank risk management, but also effectively reveal their loan gaps, So as to reduce the cost of obtaining enterprise information, provide decision-making basis for market participants, and guide investment in small and micro enterprises (Kirschenmann, 2016； Xia, 2019).

1.1.3Research object

Enterprise credit evaluation often refers to the comprehensive analysis of various factors that affect the credit willingness and credit ability of the evaluated object by the credit bureau, and the objective and fair evaluation of the performance of the enterprise in the future period. At the same time, through the construction of a scientific indicator system, the credit status of the evaluated object is expressed in the form of comprehensive scoring, credit rating, etc. (Lando, 2009). Credit evaluation is not only a "amulet" and "passport" for enterprises in the financing market for their own credit status, but also helps enterprises strengthen credit management and improve their own operations.

As this research focuses on the credit evaluation of small and micro enterprises, considering the low degree of standardization of financial management and corporate governance of small and micro enterprises, there is a lack of enough information for credit evaluation. And considering that the credit evaluation itself is the judgment of the future performance of enterprises and the degree of performance, this research uses loan data and the credit default situation of small and micro enterprises to build a credit default risk evaluation model, which is used to refer to the credit evaluation model, as the main research object of this study.

1.1.2 Significance

Relying solely on itself to accumulate funds is slow and limited in scale. The survival and development of small and micro enterprises need effective external financial support. Whether they can obtain external financial support in time and at a low price often becomes the key factor to determine the success or failure of small and micro enterprises. Under China's current financial system, Internet Finance and private lending meet the timeliness requirements and flexible lending, but the financing cost is high. Although the interest of bank loans is relatively low, the access threshold is high. Although it is the first choice for small and micro enterprises to strive for external financing, in fact, few small and micro enterprises can pass the bank's credit review. Therefore, starting with the bank's credit evaluation model, this research discusses how to help small and micro enterprises overcome the bottleneck of loan financing in commercial banks under the guarantee of bank risk control. At present, the credit evaluation model of commercial banks is mainly built based on large and medium-sized enterprises, which is not suitable for small and micro enterprises. Therefore, this research has both a theoretical and practical contribution.

(1) Theoretical significance of the study

At present, indirect financing is still the main part of China's financial system. For small and micro enterprises, credit financing is still the most cost-effective of many external financing methods, accounting for the highest proportion of all external financing. Most of the academic research on solving the financing difficulties of small and micro enterprises is still focused on the macro level, and there are still few researches focusing on the credit financing mechanism of small and micro enterprises. Thisresearch comprehensively uses bank credit theory, machine learning and other methods to build a credit evaluation model, which has important theoretical significance for in-depth understanding of the financing difficulties of small and micro enterprises.

(2) Practical significance of research

At present, small and micro enterprises constitute the main part of China's market, and have a good development prospect. However, small and micro enterprises have a short life cycle, great development pressure, and the difficulty of financing is still the biggest problem faced by small and micro enterprises. Therefore, to study the difficulty of credit for small and micro enterprises, we should start not only from small and micro enterprises, but also from commercial banks, and understand why banks are "reluctant to lend" from the perspective of commercial banks. Based on the data of small and micro enterprises, this research constructs the credit evaluation model of small and micro enterprises, which has important practical significance for further optimizing the allocation of credit resources and promoting commercial banks to better serve the real economy.

1.2 Research contents and methods

1.2.1 Research contents

On the basis of summarizing the existing research results, this study constructs the credit evaluation model of small and micro enterprises by using the credit data of small and micro enterprises, so as to provide some reference for promoting the development of credit business of small and micro enterprises of commercial banks.

The full text is divided into five chapters. The main contents of each chapter are as follows:

Chapter 1: Introduction. At the beginning of this chapter, it introduces the background, purpose and significance of the topic, expounds the research ideas, methods and main contents of this research, and finally expounds the innovation of this research.

Chapter 2: Literature review. This chapter expounds the concept of small and micro enterprises and the causes of credit difficulties, and points out that from the perspective of commercial banks, the credit difficulties of small and micro enterprises are mainly due to the difficulty in evaluating their credit level. Finally, it introduces the current main credit evaluation models, and focuses on various machine learning models.

Chapter 3: Data and methods. This chapter introduces the data and research methods used in this research, explains the potential problems of datasets and their improvement schemes, and points out that the label error of unbalanced datasets is the main problem to be solved in the empirical part of this research.

Chapter 4: Model results. This chapter is the main part of this. It shows the fitting effects of various machine learning models used, and compares the fitting results of various models.

Chapter 5: Summary, reviews and summarizes the full text research, and points out the main contribution and future improvement direction of this study.

1.2.2 Methods

This thesis starts from the general theory of the current credit default, deduces heterogeneity factors and credit environment factors which may affect SMEs' ability and willingness of reimbursement of all borrowers, and selects the corresponding proxy index.

As we all know, it is very difficult to collect information about SMEs. Since most loans are audited by local financial institutions, the content and format of the information collected vary greatly. In fact, most of the indicators used in recent literature on credit evaluation of

SMEs in China are different. We collect the datasets and indicators used in the current literature, and try to construct indicators that are similar to the previous literature used or unique based on our datasets.

The second step is to model the default samples through various machine learning methods and compare the effects of various models to select the most applicable model. This thesis reviews the most widely used machine learning models, including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, KNN, SVM and BP Neural Network, ADABOOST, XGBOOST.

Moreover, we also consider the problems of unbalanced class and of label credibility. We test several methods that are often used to improve the effects of models with datasets with unbalanced class and non-credible label.

Finally, according to the applicable model obtained in the previous step, the selected final indicators are more in line with the actual default characteristics of loans for small and micro enterprises in China, and have a stronger correlation with default probability and default loss rate.

1.3 Contribution

The existing literature about enterprises' credit evaluation, most of which refers to general enterprises, especially refers to the public company with data acquired easily, with relatively few researches focusing on the credit evaluation of SMEs. The mainly reason for few researches focusing on SMEs is that it is hard to get enough samples of SMEs with the necessary variables which is caused by the situation that accounting system and public disclosure is not necessary for SMEs.

This research summarizes the developing track of indicator system and applicable conditions of exact models of the literature on SMEs credit evaluation. The indicator system ranges from financial variables to non-financial indicators and industry characteristic indicator, with evaluating methods gradually shifting from traditional qualitative analysis to quantitative analysis, then gradually into machine-learning models.Several machine learning models are widely compared, and the importance of features is further analyzed. It provides some empirical basis for further research in this field.

Chapter 2: Literature Review

At present, the literature on the credit evaluation of SMEs is mainly divided into two aspects: one is to build the credit indicator system for SMEs and the other is the model approach(Altman, 1968; Altman et al., 1977; McCulloch & Rossi, 1994; Yeh et.al, 2012; S.W. Li & Tong,2015).

2.1 The core issue of financing for SMEs

2.1.1 Small and micro-sized enterprises

M. X. Gao et al. (2022)conclude that small and micro enterprises include four types:micro enterprises, small enterprises, family workshop enterprises and individual businesses. The division criteria of small and micro enterprises need to consider the characteristics of their industry, operating revenue, number of employees, total assets and specific departments.

The early classification standards of enterprises in China mainly classified industrial enterprises according to their production capacity, without considering the actual scale of enterprises and other industries. Dang et al. (2018) pointed out that the classification of enterprise scale is mainly carried out from both qualitative and quantitative perspectives. From a qualitative point of view, the main statement is the business scale and market position of the enterprise. Small enterprises usually refer to the enterprise, which is usually operated independently and has a small business scale, occupies a subordinate position in its industry, has a low market share and does not have market power. From a quantitative point of view, it mainly classifies the enterprise scale by setting quantitative standards, usually based on single or several indicators such as the number of employees, total assets and sales revenue. The applicable standard values are also different for enterprises in different industries. The National Bureau of Statistics(2003) issued the Categorization for Large, Medium and Small Enterprises in Statistics (Provisional), which divided industries such as industry, transportation, wholesale and retail into three categories: large enterprises, medium-sized enterprises and small enterprises, in terms of asset scale, sales and number of employees. Take an industrial enterprise as an example. If the number of employees is less than 300, or its sales is less than 30 million, or its total capital is less than 400 million, it would be classified into small enterprises.

With the development of China's economy, the main structure of the market has also undergone great changes. Compared with the internationally accepted standards, this enterprise scale division not only does not distinguish small enterprises from micro enterprises, but also its division standard is obviously too large. Take the EU as an example, as shown in the Table 2.1 below:

Table 2.1EU SME standards

Type	People	Relationship	Turnover（Million Euro）	Relationship	Asset（Million Euro）
Middle	<250	and	<=50	Or	<=43
Small	<50	and	<=10	Or	<=10
Micro	<10	and	<=2	Or	<=2

Source:M. X. Gao et al. (2022)

Too careless division of enterprise scale will lead to a large number of small and micro enterprises with relatively small scale and large-scale small enterprises being divided into the same group. Therefore, small and micro enterprises with relatively small scale will not enjoy more preferential treatment in terms of credit resources and financial subsidies, resulting in the internal differentiation of small enterprises.

The original enterprise division standard can not meet the needs of development. In 2011, the Ministry of industry and Information Technology, the National Bureau of Statistics and the National Development and Reform Commission jointly issued the Classification Standards for Small and Medium-sized Enterprises, which further subdivided micro enterprises on the basis of the original classification of small and medium-sized enterprises. Take industrial enterprises as an example: medium-sized enterprises with 300 employees or more and an operating income of 20 million yuan or more; small enterprises with 20 or more employees and an operating income of 3 million yuan or more;those with less than 20 employees or an operating income of less than 3 million yuan are micro enterprises.

The new classification standard mainly has the following characteristics: firstly, the new standard defines the scope of micro enterprises, which helps the most vulnerable enterprises benefit from government policies; secondly, the new standard refers to international practice, simplifies the division standard of small and micro enterprises, and simplifies the classification indicators from the original three to two or one, which is more conducive to the development of small and micro enterprises; thirdly, the coverage of the new standard is also more extensive, covering all national economic industry classifications.

2.1.2 The core issue

At present, small and micro enterprises have a relatively simple capital structure, a high proportion of internal capital and a low proportion of external financing. According to institutional survey data, self-owned capital and retained earnings provide more than half of the financial support for enterprises. Among them, self-owned capital accounts for about 30% of the capital of small and micro enterprises, and retained earnings accounts for 26%. However, for external financing capital, small and micro enterprises have difficulty in getting financing, issuing corporate bonds accounting for only about 1% of the total financing capital.

Based on the current credit financing research for small and micro enterprise in our country, we find that financial institutions cannot meet the demand of small and micro enterprise financing, owing toimperfection of financial market system.External channel of financing for small and micro enterprise is limited, financial services supporting measures are not sound. The main reason for the difficulty in credit financing of small and micro enterprises is the lack of supporting services such as laws and regulations, policy supervision, guarantee and credit investigation.

Compared to large and medium-sized enterprises, small and micro enterprises have small scale and low risk resistance ability (Peng, 2012). Especially in economic fluctuations, the default rate will raise sharply, which makes little contribution to the profits of banks. Therefore, credit business of small and micro enterprises is not attractive enough for commercial banks that aim tomaximize profit.

At present, the main customers of domestic commercial banks are large and medium-sized enterprises, and the same for supporting credit examination and approval mechanism. On the one hand, credit examination is mainly based on normative financial reports, audit reports and other data. On the other hand, credit examination attaches too much importance to the credit enhancement measures of enterprises and has strict requirements on collateral (L. J. Gao, 2012).

The actual situation is that small micro enterprise is at the early stage of development, and management is not mature, cannot provide standard bank financial data. Small scale, low mortgage rates and assets factors also make small and micro enterprises unable to provide full specified amount of land, real estate, machinery and equipment (Peng, 2012).

In addition, the bank credit approval process, guarantees and registration are pretty complex. This makes thatthe input and output of commercial banks' financing to small and micro enterprises are not proportional.The strict credit audit system of commercial banks and the limited human and material resources of enterprises make small and micro enterprises ‘dare not borrow’ (Peng, 2012).

In addition, the remaining external financing channels of small and micro-sized enterprises, such as issuing bonds and stocks, have not been fully developed. And there is a huge gap in relevant markets.At present, our country multi-level capital market construction is not perfect, although the Growth Enterprise Market（GEM) whose listing requirements is more loose than the main board market has launched (Peng, 2012) . In fact, enterprises which are successfully listing on the GEM already have mature financing channels and GEM give priority to the largest assets medium enterprise. The opportunity of small micro enterprises to obtain financing from GEM is extremely low. In addition, it is difficult for small and micro enterprises to issue corporate bonds based on their own credit. Even if they are successful, the cost of issuing bonds is higher than that of bank loans. Therefore, external financing channels are limited for small and micro enterprises except bank loans, especially those in the early stage of development.

The service of social credit rating agencies for small and micro enterprises is not in place, leading to the difficulty in searching credit data for small and micro enterprises. When financial institutions audit small and micro enterprises, they need to collect credit information. They often need to involve many government and non-government departments, such as industry and commerce, taxation, customs, court, etc., and this causes a credit investigation time-consuming, with high cost and incomplete data problems. So the lack of credit data seriously restricted the financial institutions to small micro enterprise credit business support.

The guarantee system of small and micro enterprises is not perfect. First of all, the service of guarantee institutions is not in place, which limits thedevelopment of credit financing business of small and micro enterprises. On the one hand, policy-oriented guarantee institutions lack market dominance and lack sufficient support for small and micro enterprises. On the other hand, the commercial guarantee system is not mature, the management is uneven. Secondly, the guarantee industry laws and regulations are not perfect, which are restricting the development of the guarantee industry of small and micro enterprises. Thirdly, the guarantee institutions of small and micro enterprises lack effective supervision. The regulators of credit guarantee institutions in some provinces and regions are scattered, and there is a lack of unified supervision subject. Finally, guarantee institutions bear large risks for the guarantee of small and micro enterprises. But policy-based re-guarantee institutions' relevant services are not in place, resulting in the difficulty of risk transfer for guarantee companies and being lack of motivation for loaning to small and micro enterprises.

The construction of laws and regulations related to small and micro enterprises is not yet perfect (L. J. Gao, 2012). Although China has implemented the ‘Law of the People's Republic of China on the Promotion of Small and Medium-sized Enterprises’ for twenty years, it lacks relevant detailed rules and supporting administrative supervision system, which cannot play a sufficient guiding role in many key issues related to the credit business of small and micro enterprises.

For an example, credit guarantee, administrative supervision, and corporate bond issuance or management have not yet issued specific rules and supporting laws or regulations with strong operability. On the other hand, some laws and regulations currently in force have negative guiding clauses for loan guarantee of small and micro enterprises. For example, ‘Guarantee Law’ and ‘General Rules for Loans’ explicitly stipulate ‘strictly control credit loans, actively promote guaranteed loans, and require proof of pledge’.

Meanwhile, fiscal and tax regulatory policies are not enough to support the financial business of small and micro enterprises. For an example, China promulgated and implemented the ‘Notice of the Ministry of Finance and the State Administration of Taxation on the Policy of Pre-tax Deduction of Agricultural Loans for Financial Enterprises and Loan Loss Reserves for Small, Medium and Micro-sized Enterprises’ (hereinafter referred to as the Notice) in 2009. The Notice provides a certain degree of tax preference for financial institutions that specialize in small and micro enterprises' credit business, and their loan loss reserves can be deducted before tax. However, the preferential term stipulated in the Notice is only two years, which weakens the policy's support for the financial business of small and micro enterprises. Secondly, the development time of small and micro enterprises is short, the management is not mature enough, and the risk resistance ability is weak (Peng, 2012). The financial institutions engaged in the credit business of small and micro enterprises are faced with high credit risk, difficult credit audit, and high cost. However, the relevant regulatory authorities have not implemented differentiated regulatory measures in capital adequacy ratio, tolerance of non-satisfactory ratio, duty exemption and other aspects.

In addition, the risk compensation fund system for the credit business of small and micro enterprises has not been effectively established. Compared with large and medium-sized enterprises, small and micro enterprises have relatively weak awareness of transferring financial risks through property insurance and other means. When risks occur, they lack corresponding risk loss compensation measures, which further weakens their ability to resist losses.

Finally, the proportion of small and micro enterprises in China actively hiring external professional consulting and management agencies to optimize enterprise management is too low. Data shows that financial and management consulting services for small and micro enterprises are not in place, the number of consulting and management agencies is small. There is still a huge market gap to improve.

2.2 Indicator system of credit evaluation for SMEs

The traditional indicator system of the credit evaluation focuses on financial indicators to examine the credit status of enterprises. However, SMEs have their own characteristics different from large enterprises. Scholars have gradually realized it and start to consider operating characteristics of SMEs, adding innovative indicators such as enterprises' growth, innovation and industrial environment to the indicator system (B. L. Fan& Zhu, 2003).

Some other scholars believe that the results of credit conditions in different industries are different(Niu, 2005). Therefore, characteristic indicators of industries are introduced to evaluate the credit status of SMEs more carefully.

2.2.1 Current situation

First, the credit rating index system of typical international rating agencies.At present, the typical international rating agencies mainly include Moody, Dun & Bradstreet and Standard & Poor's.Moody mainly considers the industry development situation, macro policy situation, management quality, company operating prospects, national regulatory environment and other specified indicators, as well as the income situation, asset efficiency, cash flow, debt operating ratio and other quantitative indicators. It evaluates the credit status of loan enterprises through the analysis and judgment of the internal laws of various rating indicators(W. Sun&Wang, 2012).

Dun & Bradstreet conducts credit rating on enterprises from business information, geographic information and other operating factors, financial factors such as quick ratio, current ratio, debt-to-equity ratio, and payment information, public record and other debt-paying willingness.

Standard & Poor's conducts credit rating from the country risk, industry characteristics, product and market conditions, strategy and management ability and other business operating conditions, as well as liquidity, cash flow adequacy, capital structure, accounting risk and other aspects of the enterprise internal financial factors(Shen, 2011).

Second, the credit rating index system of typical domestic rating agencies. Typical domestic rating agencies mainly include China Chengxin International, DagongInternational .

The credit rating index system of China Credit International focuses on the analysis of the debt paying ability of enterprises and the cash flow status. It mainly gives credit ratings to customers from six aspects: macroeconomic analysis, governance level, corporate structure, assessed debt structure, operation and financial status, and industry and regulatory trend analysis.

Dagong mainly focus on the corporate finance condition,the macroeconomic environment, policy and regulatory measures, industry development trends and other enterprise external environment factors to conduct customer credit rating(Z.J, Li, 2017; Qiu&Chen, 2014).

Third, the credit rating index system of domestic and foreign commercial banks. Citibank mainly evaluates the credit rating of enterprises from the aspects of enterprise management, competitive position, financial status and industry conditions. Industrial and Commercial Bank of China mainly focuses on economic environment, policy support and credit environment factors such as macro conditions, industry ranking, enterprise development prospect of factors such as profitability, solvency factors, and paid-in capital, guarantee ability, and shareholders to carry on the small business credit rating.

China Construction Bank, mainly from the enterprise profit ability, debt paying ability, operation ability, growth ability and other financial factors, and tax records, bank account behavior factors as well as industry characteristics, enterprise scale,carries on the small business credit rating.

2.2.2 Financial indicator

The credit evaluation of SMEs focuses on the financial status and business situation of SMEs. The traditional indicator system of credit evaluation examines the credit status of SMEs by exploring financial indicators such as solvency, cash flow, profitability and operational capacity. Among these indicators,solvency reflects enterprises' ability to liquidate its assets;cash flow is the basis of whether the enterprise can repay the money on time;the profitability reflects business condition, and operating capacity embodies enterprises' management ability.

Tan et al.(2009) selected 15 financial indicators from the above-mentioned indicators category to examine the creditworthiness of listed SMEs by Factor Analysis. W. Sun and Wang (2012) set indicators in terms of operating capacity, solvency and profitability, and then established the credit assessment model for SMEs.

2.2.3 Innovative indicator

SMEs differ from large enterprises in that they are small in size, lack of standardization, but have great potential. Many scholars consider the business characteristics of SMEs and introduce some innovative indicators in their credit evaluation research, focusing on such indicators as enterprise growth, innovation and industry environment of enterprises, to improve the overall quality of the enterprise and the quality of managers, so that it is possible to assess SMEs' credit in a more comprehensive way.

Based on financial indicators, innovation indicator, enterprise growth, development ability, industry growth and macro-environmental indicators were added by Niu(2005)to reflect SMEs' prospects and credit status in an integrated manner. Scholars such as B. L. Fan and Zhu (2003)、Qiu and Chen (2014) proposed that non-financial indicators such as innovation and growth of SMEs should be increased to measure the development potential of SMEs and stressed that the differentiated credit assessment of SMEs should pay attention to the characteristics of the whole industry, with due regard of China's national conditions and capital market development, and establish good relationship between banks and enterprises and checkout the rationality of the rating indicators timely.

H.Zhang (2008) believes that Chinese SMEs are mostly composed of private enterprises, in which managers play a decisive role in their operation and development. Therefore, it is necessary to examine the competences of managers and the comprehensive competences of SMEs.

2.2.4 Characteristic indicator

Supply chain finance is a process of optimizing the availability and cost of funds in the supply chain dominated by core enterprises(Lamoureux & Evans, 2011). In essence, it is based on the real transactions in the supply chain, and designs a series of financing schemes to solve the short-term financing problems of various enterprises in the supply chain.Supply chain finance plays an important role in the financing of small and micro enterprises.The credit evaluation based on the supply chain is no longer to examine the credit status of a single SME, but the comprehensive situation of the whole supply chain, which can effectively alleviate the embarrassing situation of SMEs.

H. Zhang (2008), from the point of view of supply chain finance, paid attention to the credit guarantee from core enterprises to SMEs, so that the financial relationship between the SME and the core enterprise should be emphasized in the selection of indicators. In the same vein, Yang et al. (2011) consider that it is necessary to increase the evaluation indicator of the upstream-downstream relationship of enterprises, analyze the situation of small and micro-sized enterprises from the perspective of the industrial chain, and focus on the concentration of upstream suppliers and downstream customers of SMEs.

There are also scholars who carry on the credit evaluation on SMEs from the perspective of Internet finance. Zheng(2015) refers that e-commerce is the data source of Internet finance which is conducive to addressing the information asymmetry of technology-based SMEs in their financing process. The asymmetry in turn leads to the demand of enterprise credit services, adding internet finance indicators such as testimonials, advance compensation, online orders share in the indicator system，which can be helpful to inspect the credit status of enterprises more carefully.

2.2.5 Feature filtering

To sum up, it can be found that there are many indicators for credit evaluation of small and micro enterprises, covering finance, macro-economy, industry characteristics and many other aspects. Although, these indicators are often highly relevant, direct adoption will not ensure accuracy, which means that we need to screen these indicators.

The traditional feature filtering methods are statistical methods. Z. J. Li (2017) selected the indicators with significant difference between the median of defaulting customers and non-defaulting customers through Brown-Mood test, and through Moses variance test, and deleted the indicators reflecting information repetition through Kendall rank correlation analysis.

B. Meng et al.(2014),through the method of combining rank correlation and rank sum test to delete the index of information duplication, established a credit rating index system of SMEs with 22 indexes including gross profit margin and the number of contract defaults was finally constructed.

Chi and Li(2019)deleted the indicators with collinearity by using the variance expansion factor, and deleted the indicators with a significance level greater than or equal to 0.05 by using the Logistic model. Finally, they selected 13 credit rating indicators that can significantly distinguish good customers from bad customers to build a credit evaluation system for small and micro enterprises.

Nikolic et al.(2013) first divided 350 indicators into 24 categories through cluster analysis and selected the indexes with the largest Implied Volatility (IV) value in each category, a total of 24 indicators. Then, 8 credit evaluation indicators were selected from the 24 indicators through Logistic regression model with the maximum Gini coefficient as the standard.

Another important indicator screening method is machine learning. In general, principal component analysis (PCA), factor analysis and cluster analysis are based on the idea of dimensionality reduction(Van Der Maatenet al., 2009). A limited number of variables are selected from a large number of candidate variables, and the selected variables represent the highest correlation with the information dimension to which they belong, so as to convert multiple indicators into fewer irrelevant comprehensive indicators, and contain most of the original information as much as possible.Machine learning not only provides a new analysis tool, but also can solve the problems of low data quality and different data dimensions that cannot be solved by traditional statistical methods.

J. Wang et al. (2012) proposed a new method, Rough Set based Feature Selection (RSFS), which combines rough set and decentralized search algorithm, to select credit evaluation indicators. The empirical results of credit database samples from Australia and Japan show that the method RSFS has advantages in saving calculation cost and improving classification performance.

Kruppa et al.(2013), through random forest method and nearest neighbor method, take 64 to 524 customers who buy household appliances by stages as samples. They select the customer's location, age, etc., seven indicators that play a key role in credit rating, so as to successfully construct a consumer credit evaluation model. OreskiandOreski(2014) proposed a heuristic algorithm combining neural network and hybrid genetic algorithm (HGA-NN) to carry out credit evaluation index selection.

To sum up, although the research on the selection of credit evaluation indicators has made important progress in academic circles at home and abroad, there are still many problems. For example, we have pointed out in section 2.2.1 that it is not only difficult to accurately identify the risk of breach of contract, but also requires a lot of manpower to evaluate the credit status of small and micro enterprises only through financial indicators. In the next section, we will briefly introduce and comment the existing credit evaluation models of small and micro enterprises.

2.3 Credit evaluation model of SMEs

The development of traditional credit evaluation model of SMEs can be divided into two stages. The first stage is the expert evaluation method which is mainly qualitative, and the second stage is the comprehensive evaluation method with quantitative factors.

The expert evaluation method is represented by the 5C evaluation method proposed by the early western commercial banks, which carries out credit evaluation from the five aspects of character, capital, capacity, collateral and condition (Abrahams & Zhang, 2008). According to their professional quality, the evaluation experts score the enterprise credit from these five aspects and add up the enterprise credit level. Character refers to the character of the borrower, that isthe repayment intention, the higher the repayment intention, the lower the possibility of default of the company will be. Capital refers to capital adequacy, which refers to the proportion of shareholders' equity investment in the total capital of the enterprise, the more equity investment accounts for, the stronger the anti-risk ability of the enterprise itself will be. If the debt accounts for too much, it will generate a lot of interest and occupy the cash flow of the enterprise, which may increase the probability of default. Capability refers to the solvency of the enterprise, which is mainly reflected in the return volatility of the borrower. If the borrower's income is unstable, its ability to repay will be limited. Collateral is the guarantee. For the guaranteed enterprise, the loss caused by the enterprise's default will be small. Condition refers to the business cycle. In reality, the rating of financial intermediaries on enterprises will be adjusted according to the business cycle.

The comprehensive rating method is an extension of the 5C evaluation method, in which indicators are given by different weights according to its relative importance, and the credit score of an enterprise is finally obtained by summing the weights of its indicators. This approach is still common now, such as analytic hierarchy process (AHP) and fuzzy comprehensive evaluation, which is developed on the basis of AHP.

The analytic hierarchy process (AHP) was proposed by American operational research scientist T.L. Saaty(2004). It is a systematic and hierarchical analysis method combining qualitative and quantitative analysis. Firstly, the hierarchical structure model is established. Based on the in-depth analysis of the problem, the relevant influencing factors are decomposed into several levels from top to bottom according to different attributes.Secondly, the paired comparison matrix is constructed, the weight vector is calculated and the consistency is tested. This method solves the complex and difficult decision-making problem by measuring the relative importance of quantitative and qualitative indicators.Huo(2012) used the factor analysis method to eliminate the correlation of these selected indicators and then used the AHP method to determine the credit indicator weight of SMEs, of which the predicted results were basically in line with the actual situation.

Much literature set the weights of indicators by analytic hierarchy process, and then introduced membership function to construct multi-scale model and comprehensive evaluation model for SMEs credit rating (Cai&Yuan, 2005; M. Fan et al., 2010; Jing&Wang, 2013), which made up for the disadvantages of evaluating the credit level of SMEs by the standards of large enterprises to some extent. Zhu et al. (2015) and L. M. Wang et al. (2016) use AHP and information entropy to establish the comprehensive weight of evaluation criteria subjectively and objectively, which overcame the problems of subjectivity, complexity and ambiguity.

Another type of credit evaluation model of small and micro enterprises is mathematical model, mainly including discriminant analysis, linear probability model, probit model and KMV model.

The discriminant analysis method emerged early and has developed many branches, one of which is the Z-score model (Altman, 1968). It is a risk-warning model for credit risk assessment of enterprises based on pure financial data. Z-score model express as equation :

$$Z\_score = \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + \beta_5 E \tag{2.1}$$

where *A* is working capital divided by total assets, *B* is retained earnings divided by total assets, *C* is earnings before interest and tax divided by total assets, *D* is market value of equity divided by total liabilities, *E* is sales divided by total assets. Later, Altman et al.(1977) improved the Z-score model from five previous variables to seven, known as the Zeta Model, which can more accurately identify companies about to go bankrupt, and can be applicable to more industries.

X. Wang et al. (2021) pointed out two defects of discriminant analysis based on the comparative analysis of different evaluation methods: first, it requires the data to obey the normal distribution, but most of the actual financial data do not obey the normal distribution;second, these two models are fitted according to the data of American enterprises, and most scholars believe that they are not suitable for small and micro enterprises in China.

Based on the above reasons, more and more scholars turn to linear probability model to explore the linear regression relationship between credit status and factors affecting enterprise credit(McCulloch & Rossi, 1994;Qing & Xin, 2015). Although the linear probability model is simple to set, it has many defects. Firstly, the linear probability model also requires the independent variables to meet the assumption of normal distribution. Secondly, the fitting value of linear probability model is often greater than 1 or less than 0, which does not accord with the definition of probability; Finally, the marginal effect of linear probability model is constant, which is not only inconsistent with reality, but also difficult to explain its economic meaning.

Due to the defects of linear probability model, discrete selection model is widely used in academia. Binary discrete selection models include logit model and probit model. The basic principles of these two models are similar. They assign probability to a cumulative distribution function to keep the estimated value between 0 and 1. At present, logit model is more and more widely used in machine learning.

The core idea of KMV model is the option pricing theory of Black and Scholes(1973) and the corporate debt pricing theory of Merton(1974). KMV company believes that the equity value of the company is essentially the same as the call option, so it extends this idea to the company's credit risk assessment, develops the KMV model, and uses the Black Scholes option pricing model to estimate the market value and volatility of the company's assets according to the market value and volatility of the company's equity, debt value and risk-free interest rate. Then it calculates and evaluates the default risk of listed companies according to the relationship between company assets and liabilities (Crosbie & Bohn, 2003). This model has been widely used (L. Zhang et al., 2004).

Domestic research on KMV model can be divided into two main aspects: first, the applicability of KMV model in China; second, further empirical research is carried out by modifying the variables in the KMV model(Yeh et al., 2012). Because KMV model is only applicable to the credit risk assessment of listed enterprises, while small and micro enterprises lack complete financial data and the data quality is not high, the promotion of KMV model in China's small and micro enterprise market needs further exploration and correction.

2.4 Machine learning model

Based on the different development stages of small and micro enterprises and their corresponding data quality, the formation and development of their credit evaluation methods have experienced three development stages: traditional model, mathematical statistics model and machine learning model. Gradually changing from subjective analysis to objective analysis, and developing from qualitative analysis to quantitative analysis, these models improve the accuracy and stability of credit evaluation.

The traditional credit evaluation method is to subjectively evaluate the reference factors of credit evaluation by special rating personnel, which is uncertain. With the gradual completion of small and micro enterprise data, credit rating methods relying on mathematical models have gradually developed. The credit evaluation models of small and micro enterprises based on mathematical statistical models mainly include discriminant analysis, linear model, discrete selection model and KMV model. These models are simple and easy to operate(Altman, 1968; Altman, 1977; S.W. Li et al., 2015; McCulloch, 1994; Yeh et al., 2012), but they can not reasonably explain their economic significance. At the same time, the statistical model has high requirements for the quality of sample data and needs to strictly obey multiple assumptions.

In recent years, machine learning is widely used in credit rating models. Compared with mathematical statistical models, machine learning methods have lower requirements for data quality and more assumptions to obey. The main methods of machine learning, such as decision tree model(Freund, 1999), support vector machine(Suykens & Vandewalle, 1999), BP neural network(Jin et al., 2000) and random forest(Liaw & Wiener, 2002), have been widely used in credit evaluation research (Marqués et al., 2012).

Qing andXin(2015)established the credit evaluation model of listed small and medium-sized enterprises based on logistic regression. It's empirical results show that the overall prediction accuracy is 65%. Yi(2007)established a decision tree model to evaluate the credit of small and micro enterprises, and its prediction accuracy is more than 85%.

Tan et al.(2009) constructed a three-layer BP neural network model for credit evaluation indicators of small and micro enterprises. It is found that the credit status of small and micro enterprises is highly heterogeneous, and the credit evaluation model based on BP neural network can effectively alleviate the adverse impact of heterogeneity on the estimation results. Chang(2015), using BP neural network model to debug and calculate the credit rating index system of small and micro enterprises, found that the credit rating model of small and micro enterprises established by BP neural network has higher accuracy and better robustness than the credit rating model established by general index system. W. Chen(2012) empirically analyzed and compared the practical application of SVM and BP neural network in the credit rating model of small and micro enterprises. The results show that the accuracy and robustness of SVM are better than BP neural network model.

The above research shows that the performance of machine learning model in credit evaluation of small and micro enterprises is better than that of traditional evaluation methods and mathematical statistical models, but there are also great differences in the application effects of different machine learning models. This section will summarize the advantages and disadvantages of machine learning models

2.4.1 Logistic model

2.4.1.1 Brief introduction

Logistic model predicts the credit risk based on the regression analysis of the existing sample data, and then sets the credit risk warning line according to the risk preference of different users for defining a risk threshold (McCulloch & Rossi, 1994). For a sample*i*, since the value of the explained variable Y_i is 0 or 1, we can regard Y_i as the realization value of the random variable Y_i : the probability of Y_i taking 1 is π_i, and the probability of taking 0 is 1- π_i. The random variable Y_i follows the (0-1) distribution with parameter π_i, and the distribution law of$Y_i$isshown in equation :

$$\mathbf{Pr}[Y_i = y_i] = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.2}$$

Obviously, if $Y_i = 1$, the probability of Y_i is π_i,; If $Y_i = 0$, the probability of Y_i is 1- π_i.

Peng (2012) used principal component analysis to screen out public financial indicators with good independence, and added common factor and default distance (DD) into logistic regression analysis. The results showed that the logistic regression model with default distance had higher prediction accuracy than the ordinary logistic model.

One of the drawbacks of logistic regression model is that its decision-making process is linear, which makes logistic regression model unable to solve complex nonlinear problems, and is difficult to capture new business characteristics and new enterprise type (Klieštik et al., 2015) . This shortcoming is unable to dynamically evaluate enterprise credit, and is easy to cause credit evaluation error. And the model is very sensitive to missing values and extreme values. Too many missing or extreme values of sample data will be negative to the fitting effect of model.

2.4.1.2 Penalty

In the process of machine learning, because of the large amount of data provided for training, there will be many dimensions (e.g., variables) generated during this process. Some dimensionsare important while others are irrelevant. The more the dimensions of judgment, the worse the generalization ability of the model. Therefore, we will add an additional item into the loss function, that is the penalty item, to restrict the loss function so that the model can balance the number of dimensions and generalization ability. Penaltyitem generally includes L1-norm and L2-norm. In linear regression models, models that use L1-norm are called Lasso regression models, and models that use L2-norm are called Ridge regression models (H. Wang et al., 2015) .

The general form of the L1 penalty is expressed as equation :

$$\|w_1\| = \Sigma|w_i| \tag{2.3}$$

L1-norm refers to the sum of the absolute values of the elements in the weight vector w of the function optimization, which is usually expressed as $||w||_1$.L1 can be used for feature selection and preventing overfitting. Because of a coefficient α usually added before the regular term ,the principle is to make α less than 0, then the more zeros in w for the penalty term, the better the model needs to weigh the complexity of the model and the effect of the model in the global optimization, and the complexity of the model depends on the size of α.

The general form of the L2-norm is expressed as equation :

$$\|w_2\| = \left(\Sigma|w_i|^2\right)^{\frac{1}{2}} \tag{2.4}$$

The main role of L2 is to prevent overfitting. L2 makes the fitting process tend to make the weights as small as possible, and finally constructs a model with all the parameters smaller. Because it is generally believed that models with small parameter values can adapt to different datasets, over-fitting is avoided to some extent (for example, if the parameters of the linear regression equation are too large, as long as the data changes slightly, it will have a great impact on the prediction results.If the parameters are small enough, the data change will have little impact on the results, that is, the anti-interference of the model is very strong).

2.4.1.3 Feature importance

Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction.Feature importance scores can be calculated for problems that involve predicting a numerical value, called regression, and those problems that involve predicting a class label, called classification.

For any features, the corresponding score can be expressed as equation :

$$\mathbf{P'} = \frac{1}{1+e^{-\beta X'}} \tag{2.5}$$

where$X'$is the feature vector of the sample. Make a transformation, we can getequation :

$$J = \ln\frac{\mathbf{P'}}{1-\mathbf{P'}} = \beta X' \tag{2.6}$$

Then the contribution of the feature is equation :

$$Importance = \left|\frac{\beta_i X_i'}{\beta X'}\right| \tag{2.7}$$

The higher the contribution, the greater the impact of the indicator on risk (the target variable).

2.4.2 Decision tree model

The decision tree is a tree-structure decision-making method, which classifies source data in a tree pattern, each node of the tree representing a certain classification condition. Then the lower nodes and branches are repeatedly established in each branch sub-set to generate the decision tree. Yi (2007) established a decision tree model to assess the credit worthiness of SMEs and the prediction accuracy reached more than 85%, as well as Xu and Ma(2011).

The Decision tree model has many advantages, such as low cost of machine learning, strong ability of model interpretation and so on. Using the decision tree model for credit evaluation can easily describe and express the relationship and interaction between local and overall decisions in different stages, and accurately find the key factors of enterprise default.

2.4.2.1 Construction of a decision tree

According to Safavian &Landgrebe(1991), the process of constructing a Decision Tree is a process of splitting the sample set. On each step of construction, we actually split the sample set into several sample subsets.

The core idea of setting up this model is to figure out the order of classification on each feature attribute and the structure of the hierarchy tree. The key of constructing a decision tree with accurate prediction is to choose the feature attributes which have better distinguishing ability. The principle of evaluating the splitting of a Decision Tree model is to check whether it can classify the sample subsets of the same category which means a better classification.

The Decision Tree model is based on local optimum rather than global optimum, and it is a greedy algorithm, i.e. splitting each time according to local optimal feature attributes. There are many related algorithms to implement the Decision Tree model, and here we introduce two of the most widely used algorithms C5.0 and CART among them.

2.4.2.1.1ID3, C4.5 and C5.0 algorithm

The oldest version of C5.0 algorithm is ID3 which was developed by J. R.Quinlan(1986). ID3 algorithm was proposed mainly aimed at discrete attribute data. Later, it was continuously improved to form C4.5, which added the discretization of continuous attributes on the basis of ID3. C5.0 is a classification algorithm applied to large datasets by C4.5, which has been improved in terms of execution efficiency, memory usage and adapting multi-branching decision tree. Next, we will introduce them one by one.

J. R.Quinlan(1986, 2014)proposed a splitting method for the Decision Tree model which follows the ‘Maximum Information Entropy Gain’ principle on each step of splitting and name it as ID3 algorithm(Hssina et al., 2014). For example, if the chosen feature attribute has N possible values, then according to ID3 algorithm, the sample set will be split into N sample subsets. The gain of Information Entropy means the change of Information Entropy before and after the splitting of the sample set. The more the gain of Information Entropy is, the more accurate is our prediction from the splitting. Beyond abstract mathematic concepts, Information Entropy is often construed as the amount of information cost to eliminate some sort of uncertainty in practical life.

In other words, Information entropy indicates the amount of information contained in an uncertain event. The lower the Information Entropy is, the less will be cost to eliminate the risk of an uncertain event, which means the more accurate is the partition of sample in a binary-classification problem. In the practical implementation of ID3 algorithm, feature attributes with more possible values are often chosen to be the criteria of splitting.

However, when choosing such attributes according to the ‘Maximum Information Entropy Gain’ principle, we may be trapped into a local optimal solution rather than a global optimal one. C4.5 algorithm(Hssina et al., 2014)applies the ‘Maximum Information Entropy Gain Ratio’ principle instead of the ‘Maximum Information Entropy Gain’ principle, which can solve the ID3 algorithm’s problems. The rule of ‘Maximum Information Entropy Gain Ratio’ considers the number of sample subsets after splitting, and calculates the new information scale, i.e. computes the ratio of Information Entropy Gain to the number of new subsets produced by the splitting as the Information Entropy Ratio, and then choose the feature attribute that cause the greatest increase in Information Entropy Ratio. This method can reduce the occurrence of classifications that are hard to explain.

C5.0 is the newest version of C4.5. The major improvement is adaptability to big data and adaptability to multi-branching decision tree. C5.0 can translate continuous attributes into discrete variable, which can increase the accurancy to predict.It can also reduce the GPU occupancy rate and save runtime.Its efficiency of prediction has a range of ascension compared with the C4.5(Hssina et.al, 2014).

CART Decision Tree Model could either be Classification Tree and Regression Trees (CART algorithm used in Regression Trees can also be named C&RT). Actually, most of relative research use Classification Tree in SME credit evaluation field(Steijvers et al., 2005;Xu&Ma, 2011), and so do we in this study. So we just introduce CART algorithm. The

CART algorithm introduces Gini coefficient to measure the purity or uncertainty of data(Hssina et.al, 2014)which can be expressed as equation :

$$\boldsymbol{Gini} = \mathbf{1} - \sum p_i \tag{2.8}$$

where p_i is the proportion of the i-th sample. The less pure the subset sample, the larger the GINI value. If the subset has only one sample, the Gini coefficient reaches the maximum value. The evaluation indicators for the division effect can be expressed as equation :

$$\boldsymbol{GAIN} = \sum \boldsymbol{a}_i * \boldsymbol{Gini}_i \tag{2.9}$$

where a_i is the proportion of the i-th sample, $\boldsymbol{GINI_i}$ is the Gini coefficient of the i-th subset; taking the first split as an example, as the equation shown:

$$\boldsymbol{Gini} = \frac{\boldsymbol{N}_1}{\boldsymbol{N}_1 + \boldsymbol{N}_2}\boldsymbol{Gini}_1 + \frac{\boldsymbol{N}_2}{\boldsymbol{N}_1 + \boldsymbol{N}_2}\boldsymbol{Gini}_2 \tag{2.10}$$

Optimal segmentation point makes the node's GAIN reach the maximum value

2.4.2.2 Defects of the decision tree model

In an ideal world, the Decision Tree model is able to recursively iterate through all the samples completely and unerringly. That is to say all samples could be classified accurately by feature attributes. However, such a completely accurate classification of samples may not be found in most of the time, i.e. the optimal solution of the Decision Tree model is hard to find.

The main problems of the Decision Tree model are as follows.

First, the classification rules are complex. Local greedy algorithm is often used in decision tree generation. When splitting nodes, only one attribute is selected for analysis each time. Therefore, there is a pruning method of decision tree,and it further increases the complexity of decision tree algorithm(Hand & Henley, 1997; G. Wang et al., 2011).

Second, overfitting is an issue when training a tree, which limits the generalization ability of the model. In the process of decision tree generation, sometimes the classifier design is too complex, resulting in too many sample sets, and the noise can also adapt to the classifier, resulting in overfitting. Moreover, the decision tree model is more suitable for dealing with discrete variable data samples, and has poor adaptability to continuous variable datasets (Jadhav & Channe, 2016) .

Third, there is the possibility of local optimal solution rather than global optimal solution. There is no backtracking mechanism in the generation process of decision tree. When a certain attribute is selected for testing when the node is split, it is easy to cause the model to converge to the local optimal solution rather than the global optimal solution because there is no backtracking and retesting(Kamiran et al., 2010).

2.4.2.3 Pruning of the decision tree model

In order to solve the over fitting problem of decision tree model, some sample subsets need to be deleted after segmentation, such as deleting some feature attributes or the possible values of some feature attributes, that is, pruning the decision tree. Pruning can reduce the complexity of the decision tree and reduce the over fitting problem. The commonly used methods of pruning decision trees are divided into pre pruning and post pruning.

The construction of all decision trees will stop creating branches only when the entropy cannot be further reduced. In order to avoid overfitting, you can set a threshold, and the entropy reduction is less than this threshold. Even if entropy can be reduced furtherly, you should stop creating branches. The commonly used method for pre-pruning are as follows:

(1) Stop the splitting process when the model has split for certain times

(2) Stop the splitting when the number of samples in a sample subset is smaller than a previous set threshold.

(3) Stop the splitting when the Information Entropy Gain brought by the splitting is smaller than a previous set threshold.

However, pre-pruning is not good at reducing overfitting. Once the splitting of the Decision Tree model is stopped, turning the Nth Node to a leaf node, the following splitting to optimize the model will also be stopped accordingly, which cause a lower purity and lower accuracy of the classification model.

The post-pruning method requires the Decision Tree to grow continuously and to have enough branches as far as possible until all leaf nodes have the smallest degree of impurity. Then, the variation of the impurity after eliminating a certain leaf node is calculated and a threshold value is set. When the increase of impurity brought by eliminating a certain leaf node is less than the threshold value, it indicates that the leaf node has little influence on the splitting of the sample set and need to be pruned. This pruning process is exactly the opposite of the Decision Tree's splitting process.

The post-pruning process makes full use of all the information in the training set, but it is often accompanied by a large increase in the amount of computation. Therefore, the post-pruning method is suitable for small sample sets, while for large sample sets, the computational efficiency of post-pruning method is often unable to meet the application requirements.

The pruning process of the Decision Tree model to some extent reduces the problem of its poor generalization ability caused by over-fitting, but it cannot solve the fundamental problem. Many practices have proved that Random Forest algorithm is a feasible scheme to solve the problem of over-fitting of Decision Trees.

2.4.3 Naive bayes

2.4.3.1 Brief introduction

In reality, there is often such a kind of problem in pattern discrimination that the number of features is much greater than, or equivalent to the number of training sets and there are always correlations between these features. Such problems will make the model too complicate and be likely to cause over-fitting problems.

The Naive Bayes is a classification method based on Bayesian theorem and the assumption of independence of feature conditions. For example, if a fruit has features such as red, round, about 4 inches in diameter, etc., this fruit will be judged as an apple. Although some of these feature attributes are correlated and one may be able to be determined by others, a Naive Bayesian classifier holds the idea that all these attributes are independent on the probability distribution of determining whether this fruit is an apple.

The Naive Bayes has been widely used in credit evaluation.Ye and Lu (2017) used Naive Bayes to evaluate the credit risk with German bank credit card business dataset.

2.4.3.2 Theoretical basis

Suppose there are N possible class tick marks $\boldsymbol{y} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_N}\}$, which is the loss caused by misclassifying a sample that is actually marked as c_j into c_i. Based on the posterior probability$\boldsymbol{p(c_i|x)}$, the expected loss resulting from classifying the sample x into c_i can be obtained, is the 'conditional risk' on the sample x, which can be written as equation :

$$\mathbf{R}\left(c_i|x\right) = \sum_{j=1}^{N} \lambda_{ij} P(c_i \mid x) \tag{2.11}$$

We need find a criterion$\boldsymbol{h}$: $\boldsymbol{\chi} \longrightarrow \boldsymbol{\mathcal{Y}}$ to minimize the equation :

$$R(h) = E_x[R(h(x) \mid x)] \tag{2.12}$$

Obviously, for each sample x, if h can minimize the conditional risk $R(h(x)|x)$, then the overall risk $R(h)$ will also be minimized. This produces Bayesian criteria: to minimize the overall risk, simply select the category tag on each sample that minimizes the conditional risk $\boldsymbol{R(c|x)}$, which is the equation :

$$h^*\left(\mathrm{x}\right) = \mathbf{argmin} R(c \mid x) \tag{2.13}$$

At this time,$h^*(\mathrm{x})$ is called the Bayesian optimal classifier, and the overall risk $\boldsymbol{R}(\boldsymbol{h}^*(\boldsymbol{x})$ is called Bayesian risk $\mathbf{1}-\boldsymbol{R}\left(\boldsymbol{h}^*(\boldsymbol{x})\right)$, which reflects the best performance that a classifier can achieve, is also the theoretical upper limit of the model's accuracy that can be produced by machine learning.

Specifically, if the goal is to minimize the classification error rate, the misjudgment loss λ_{ij} can be written as equation :

$$\lambda_{ij}=\begin{cases}\mathbf{0},\boldsymbol{if}\ \boldsymbol{i}=\boldsymbol{j}\\ \mathbf{1},\boldsymbol{otherwise}\end{cases} \tag{2.14}$$

Conditional risk at this time is equation :

$$\boldsymbol{R}\left(\boldsymbol{c}|\boldsymbol{x}\right)=\mathbf{1}-\boldsymbol{P}(\boldsymbol{c}\mid\boldsymbol{x}) \tag{2.15}$$

Thus, the Bayesian optimal classifier that minimizes the classification error rate is equation :

$$\boldsymbol{h}^*\left(\boldsymbol{x}\right)=\mathbf{arg\,max}\,\boldsymbol{P}(\boldsymbol{c}\mid\boldsymbol{x}) \tag{2.16}$$

That is, for each sample x, a category flag that maximizes the posterior probability $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$ is selected.

It is not difficult to find that by using Bayesian criteria to minimize decision risk, we must first obtain the posterior probability $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$. However, this is often difficult to obtain directly in real-life tasks. From this perspective, what machine learning is to achieve is to estimate the posterior probability $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$ as accurately as possible based on a limited set of training samples. In general, there are two main strategies: given *x*, *c* can be predicted by directly modeling $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$, thus obtaining a ‘discriminant model’, or the joint probability distribution $\boldsymbol{P}(\boldsymbol{c},\boldsymbol{x})$modeling, and then $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$ is obtained, thus obtaining a ‘generating model’. For the generative model, equation must be considered:

$$\boldsymbol{P}\left(\boldsymbol{c}|\boldsymbol{x}\right)=\frac{\boldsymbol{P}\left(\boldsymbol{c},\boldsymbol{x}\right)}{\boldsymbol{P}\left(\boldsymbol{x}\right)} \tag{2.17}$$

Based on the Bayesian theorem, $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$ can be written as equation :

$$\boldsymbol{P}\left(\boldsymbol{c}|\boldsymbol{x}\right)=\frac{\boldsymbol{P}(\boldsymbol{x}\mid\boldsymbol{c})\boldsymbol{P}\left(\boldsymbol{c}\right)}{\boldsymbol{P}\left(\boldsymbol{x}\right)} \tag{2.18}$$

It is not difficult to find that the main difficulty in estimating the posterior probability $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$ based on the Bayesian formula is that the class conditional probability $\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{x})$ is the joint probability of all the attributes, which is difficult to obtain from the limited training samples to directly estimate. To circumvent this obstacle, the Naive Bayes classifier uses the ‘attribute conditional independence hypothesis’, assuming that all attributes are independent of each other for known categories. In other words, the assumption is each attribute independently affects the classification result. Therefore, for a problem of multi-attribute sample, the Bayesian formula can be written as equation :

$$P(\mathrm{c}|x_1x_2\ldots)=\frac{P(c)P(x_1x_2\ldots|c)}{P(x_1x_2\ldots)}=\frac{P(c)P(x_1|c)P(x_2|c)\ldots}{P(x_1)P(x_2)\ldots} \tag{2.19}$$

where $\boldsymbol{P(\mathrm{c}|x_1x_2\ldots)}$ means the probability of a sample's being classified into the c category, x_n means the attribute used for classifying and $\boldsymbol{P(x_n)}$ is its probability.

To sum up, the core idea of Naive Bayes is to set the attributes used for classification as mutually independent ones, calculate each of its probability and finally use them to conduct the classification.

2.4.3.3 Advantages and disadvantages

The Naive Bayesian model is based on classical mathematical theory and has a relatively strong theoretical basis, efficient at dealing with multi-classification problems(Krichene, 2017). Moreover, as it needs only a few arguments and is not sensitive to missing values, it is widely used in text classification field. Meanwhile, Naive Bayes also has some defects which make it hard to be applied in some situation:

(1) The assumption of independence of feature attributes used by the Naive Bayesian model is always unacceptable in practical application. When the correlation between attributes is large, the classification results will become worse.

(2) The values of prior probabilities have to be known. Since such probabilities are often acquired by making assumptions but there are always too many models we can choose to assume, thus under some circumstances, the prediction may not be well enough because of the prior model we assume.

2.4.4 K-nearest neighbor classification

2.4.4.1 Brief introduction

The KNN algorithm(Hwang & Wen, 1998), also known as the K Nearest Neighbor method, is a non-parametric classification technique based on analog learning, which is very effective in statistical-based pattern recognition. The KNN algorithm is a classification algorithm with supervised learning, and does not need to generate additional data to describe the rules. Its rule is the data (sample) itself. It does not require data consistency, that is, there can be noise(Wang et al., 2020),.

The KNN algorithm is a theoretically mature method, originally proposed by Hart (1968).

The idea is very simple. For samples of a certain category, k nearest neighbors are found in the training set of sample space according to the Euclidean distance, and the sample belongs to the category that contains the most samples of k nearest neighbors. The basic principles of the KNN classification algorithm are as follows.

First, the sample to be classified y is expressed as a feature vector consistent with the sample of the training sample.

Then, according to the distance function, calculate the distance between the sample y and each training sample, and select K samples with the smallest distance from the sample as the K nearest neighbors of y.

Finally, the category of y is judged based on the K nearest neighbors of y.

K-nearest neighbor can be expressed as equation :

$$\boldsymbol{y} = \mathbf{arg}\max_{c_j} \sum_{x_i \in N_k(x)} \boldsymbol{I}\left(y_i = c_j\right) \tag{2.20}$$

where c_j represent class variable, $N_k(x)$ is the set of k nearest neighbors and y_i is the class of x_i. The KNN algorithm must consider two basic factors: The number of nearest neighbor samples K and the scale of the distance. The distance scale refers to a non-negative function which is used to measure the similarity between different data. In the KNN algorithm, the best choice of the model (especially the K value) is often verified by a large number of independent test and multiple models.

2.4.4.2 Advantages and disadvantages

The KNN classification algorithm is a non-parametric based classification technique. It can reach high classification accuracy for unknown and non-normally distributed data, and has many advantages such as clear concept and easy implementation. However, there are also problems in the classification process, such as the similarity calculation is too large, the distance function is too dependent on the sample itself and the similarity of the measurement is not applicable. We summarize the advantages of the KNN classification method as follows:

- The idea is very simple and intuitive, and easy to implement.
- There is no need to generate additional data to describe the rules. Theonly rule is the training data (samples) itself. It is not a requirement for data consistency, there can be noise.
- Although the KNN algorithm relies on the limit theorem in principle, it is only related to a very small number of adjacent samples in the category decision. Therefore, this method can better avoid the imbalance of the samples size.

- The KNN algorithm makes the most use of the similarity among samples, which reduces the adverse effects of improper selection of the category features, and can minimize the error in the classification process. For some categories whose category features are not obvious, the KNN algorithm can better reflect the independence of its classification rules, making it possible to implement convenient classification self-learning.

The shortcomings of the traditional KNN algorithm mainly include the following points:

- Classificationspeed runs slowly.Nearest neighbor classifier is a lazy learning method based on instance learning, because it does not actually construct a classifier (according to the given training samples). To treat a subsample which needs to be classified, it is necessary to calculate the similarity with each sample in the training sample in order to obtain the nearest K samples. For high-dimensional samples or large sample sets, the time and space complexity is high, and the time cost is O(m*n), where m is the spatial feature dimension of the vector space model, and n is the training sample set size.
- The sample library has a strong capacity dependency.The problem of strong sample size dependence is more limited in the practical application of KNN algorithm: there are many categories that cannot provide enough training samples, so that the relatively uniform feature space conditions required by KNN algorithm cannot be satisfied, so that identification error is increased.
- The features work the same.Unlike the decision tree induction method and the neural network method, the traditional nearest neighbor classifier considers each attribute to be the same (giving the same weight). The distance of the sample is calculated from all the features (attributes) of the sample. Among these features, some are strongly related to classification, some are weakly related to classification, and some features (probably most) are not related to classification. Therefore, assigning the same weights to these features may lead to misclassification.
- Determination of K value.The KNN algorithm must specify the K value. If the K value is not properly selected, the classification accuracy cannot be guaranteed.

2.4.5 SVMmodel

2.4.5.1 Brief introduction

The SVM model maps a linearly inseparable space to a high-dimensional linearly separable space through a nonlinear transformation (kernel function) to find an optimal classification hyperplane, and correctly separates the two types of samples and maximizes the classification interval at the same time, thus reducing the classification error, which then reduces the error in classification. Computing the SVM classifier amounts to minimizing an expression of equation :

$$\left[\frac{1}{n}\sum\max_{w}\left(0,1-y_i\left(w*x_i-b\right)\right)\right]+\lambda w^2 \tag{2.21}$$

where y_i is class variable, x_i is a vector that represents sample *i*.

W. Chen (2012) made an empirical comparison between SVM l and BP neural network applied in SMEs credit evaluating model, the results of which showed that the accuracy and robustness of SVM are superior to the BP neural network model.

Xia (2013) established the SMEs credit evaluation model based on SVM and compared the classification results of different kernel functions. He found that SVM with radial basis kernel function has the best classification effect. At the same time, the author compared it with the neural network model, and found that the classification effect of the SVM model is better.

The SVM model is based on the structure risk minimum principle and the VC (Vapnik-Chervonenkis) dimension theory, which sets a balance point between complexity and learning ability aiming to solve small sample, nonlinear and high-dimensional recognition problems.

The disadvantage of SVM model is that there is no proper method to determine the kernel function which is the mapping of high-dimensional space to low-dimensional space. So for general problems, SVM only turns the complexity difficulty of high-dimensional space into the difficulty of finding kernel function. In general,it’s sensitive to missing and extreme values.

2.4.5.2 Linear and nonlinear SVM

The SVM is developed from the optimal classification hyperplane in the case of linear separability. The essence is to find the support vector with the optimal classification hyperplane in the training sample, which is mathematically reduced to a solution for inequality constraint condition in quadratic programming problem. Given a training set $\boldsymbol{\Omega}=\{(\boldsymbol{x_i},\boldsymbol{y_i})|\mathbf{i}=\mathbf{1},\mathbf{2},\ldots,\mathbf{l}\}\subset\boldsymbol{R^m}*\{\mathbf{-1},+\mathbf{1}\}$, where $\boldsymbol{x_i}\in\boldsymbol{R^m}$ is the input vector. The corresponding binary classification label is $\boldsymbol{y_i}\in\{\mathbf{-1},+\mathbf{1}\}$. Assuming that Ω is linearly separable, there is a hyperplane H which can be expressed as equation :

$$\langle w,x\rangle + b = 0 \tag{2.22}$$

where $\langle w,x\rangle$is the inner product of two vectors, w is the weight, and b is a constant .

When the two types of samples are linearly separable, the conditions are met:

$$\begin{gathered} \langle w,x_i\rangle + b \ge 1, if\ y_i = 1 \\ \langle w,x_i\rangle + b \le 1, if\ y_i = -1 \\ y_i\left[\langle w,x_i\rangle + b\right] \ge 1, if\ i = 1,2,\ldots,l \end{gathered} \tag{2.23}$$

Where y_i is the category of x_i. The equation means that distance d from the point on the sample space to the classification hyperplane, by spatially resolved geometric theory to get:

$$d = \frac{\left|\langle w,x_i\rangle + b\right|}{w} \tag{2.24}$$

And because these points are closest to the classification hyperplane, the sample points with the category +1, -1 satisfy equation :

$$w,x_i + b = \pm 1 \tag{2.25}$$

which can be written as equation :

$$d = \frac{1}{\|w\|} \tag{2.26}$$

Therefore, the interval d of the classification is maximized, even if the plane with the smallest $||w||$ is the optimal classification plane. In summary, the problem of generating the largest classification interval between sample points of categories +1 and 1 corresponds to the following optimization problem which is shown in equation :

$$\begin{gathered} \min_w \Phi(w) = \min_w \frac{1}{2}\|w\|^2 \\ \text{s.t.}\, y_i\left[w,x_i + b\right] \ge 1,\ i = 1,2,\ldots,l \end{gathered} \tag{2.27}$$

The essence of the problem is to solve the quadratic programming problem of an inequality constraint.

In the case where the sample is nonlinearly separable, the general idea of the traditional statistical method is to use a nonlinear transform $\emptyset(x)$ to map the input data to a high-dimensional feature space, then perform linear classification in the high-dimensional feature space, and finally map back the former space, becomes a nonlinear classification of the input sample.

This method will encounter the problem that the dimension of the feature space is very high, which makes the calculation cost too high or even impossible. For example, constructing a 4th or 5th order polynomial in 200-dimensional space must be constructed in hundreds of millions of dimensions.

Hyperplanes can cause dimensional disasters. The way to avoid dimensionality disasters is to use kernel functions. Constructing a class hyperplane in the feature space does not need to represent the feature space in the form of a display. We only need to calculate the inner product of the vector in the feature space. So assume that the input vector is mapped to a Hilbert space as equation shows:

$$\left(\varnothing_1(\mathrm{x}),\varnothing_2(\mathrm{x}),\ldots,\varnothing_n(\mathrm{x})\right) \tag{2.28}$$

According to Hilbert-Schmidt theory, the inner product in Hilbert space has an equivalent expression equation :

$$\boldsymbol{h}_1,\boldsymbol{h}_2=\sum_{i=1}^{\infty}\boldsymbol{a}_i\boldsymbol{h}_i\left(x_1\right)\boldsymbol{h}_i\left(x_2\right)\Leftrightarrow \boldsymbol{K}\left(x_1,x_2\right) \tag{2.29}$$

Among them, $\boldsymbol{K(x_1,x_2)}$is a symmetric function that satisfies the Mercer theorem and is called a kernel function. The basic idea of the kernel method is that for any kernel function $\boldsymbol{K(x_1,x_2)}$ that satisfies the Mercer condition, there is a feature space $\left(\boldsymbol{\emptyset_1}(\mathbf{x}),\boldsymbol{\emptyset_2}(\mathbf{x}),\ldots,\boldsymbol{\emptyset_n}(\mathbf{x})\right)$, where the kernel function in a feature space generates an inner product.

It can be seen that the inner product operation of the sample space has been replaced by a nucleus.In fact, the operation is performed in the sample space rather than in the high-dimensional feature space. This is the idea of nuclear techniques.

Since the kernel function of the input space is essentially the equivalent form of the product in the feature space. Therefore, in the actual calculation, we do not have to care about the specific form of the nonlinear mapping $\emptyset(\mathbf{x})$, only need to select the kernel function $\boldsymbol{K(x_1,x_2)}$, the kernel function is relatively simple, and the mapping function may be very complicated, and the dimension is very high.Therefore, the introduction of nuclear methods has overcome the ‘dimensional disaster’ problem.

2.4.5.3 Advantages and disadvantages

The SVM model effectively balances the relationship between modeling complexity, operational efficiency, generalization and predictive stability, and has the following advantages over other machine learning methods:

- **Adopt the principle of structural risk minimization.** The principle of structural risk minimization is an important advantage of the SVM compared with other intelligent learning models. It no longer pursues the minimization of empirical risk of sample data, but introduces the concept of confidence risk, which minimizes structural risk and greatly improves the generalization ability of the model.

- **Effectively obtain the global optimal solution.** The SVM model ingeniously transforms the quadratic programming problem under the original constraint into the dual quadratic optimization problem, and finds the global optimal solution through the operation, thus effectively avoiding the local maximum possible in other intelligent learning methods. The problem is solved, which effectively improves the prediction accuracy of the model.
- **Effectively avoid dimensional disasters.** After the transformation of input data from low-dimensional to high-dimensional, by introducing the concept of kernel function, it is possible to find and construct feature classification surface in high-dimensional space, and skillfully solve the dimensionality disaster problem that plagues traditional machine learning methods.
- **Model adaptability is good.** The SVM model needs to select the kernel function in the construction process and needs to adjust the parameters to improve the model classification effect. However, compared with the Bayesian network and the neural network model, the content that needs to be adjusted has been greatly reduced. The SVM model can learn and adapt the sample data, which effectively reduces the complexity of modeling.

At the same time, the disadvantages of SVM are:

- Each classifier is trained to use all samples as training samples, so that when solving quadratic programming problems, the training speed will decrease sharply as the number of training samples increases.
- For the case where the data of the negative class sample is much larger than the data of the positive class sample, that is, the case of sample asymmetry occurs, and the influence on the model is very large. Solving the problem of asymmetry can introduce different penalty factors, and the positive class with fewer sample points uses a larger penalty factor. Also, when new categories are added, all models need to be retrained.

2.4.6 BP neural network

2.4.6.1 Brief introduction

The BP neural network model consists of an input layer, an output layer and one or more hidden layers. The individual neurons in the same layer are independent from each other and can be regarded as a highly nonlinear mapping from input layer to output layer, composed of forward propagation and reverse propagation.

Chang(2015)debugged and calculated the credit rating index system of small and micro enterprises based on BP neural network and found that the precision and error of the SME credit rating model based on BP neural network is very stable. Furthermore, it is found that the credit score prediction accuracy and error of the model are both higher than that of predicting whether a loan occurs. It indicates that each BP neural network with different output result in the assessment of the SMEs credit status has its own strength, which can be mutually integrated and confirmed.

Tan et al. (2009) andH.Y.Zhang & Li(2017) found that the network with trainlm(Levenberg-Marquardt algorithm) as the training function has the fastest convergence speed and the smallest error no matter how the structure of the hidden layer of BP neural network changes.

Zhong & Jia(2005) built a credit evaluation model for SMEs based on BP neural network and introduced negative samples as feature samples to enhance the generalization ability of the system. Similarly, Wu (2013) argued that it is necessary to introduce negative samples as feature samples in the credit rating model of SMEs, which can enhance the generalization ability of the system.

Khashman(2010) compared the evaluation performance of the neural network credit evaluation model with different topology structures and learning plans (the proportion of data used for training and verification), and found that the simpler the topology structure is, the shorter time the operation will take. When the ratio of training and validation is 40%: 60%, the model performance is optimal.

2.4.6.2 Algorithm

The predecessor of the BP neural network is the feedforward neural network. The feedforward neural network uses a unidirectional multi-layer structure in which each layer contains several neurons, and the neurons in the same layer are not connected to each other, and the information transmission between layers is performed in one direction.

The single-layer feedforward neural network is the simplest artificial neural network. There is only one output layer. The value of the output layer node (i. e., the output value) is directly obtained by inputting the value and the set weight value. A single-layer neural network can classify information into equation :

$$s_j = \sum_{i=1}^{n} \omega_{ji} x_i - \theta_j$$
$$y_j = f\left(s_j\right) = \begin{cases} 1, s_j \geq 0 \\ 0, s_j \leq 0 \end{cases} \tag{2.30}$$

The information received by the output layer is the input value s_j after the linear transformation of the input x_i, $f(\cdot)$ is the activation function, and the s_j can be classified by the excitation function.

A multi-layer feedforward neural network consists of an input layer, one or more hidden layers, and an output layer. Each layer is equivalent to a single-layer feedforward neural network. The input is the output of the previous layer (the input layer input is the external information input), and then the output is output as the next input until the last layer. Multi-layered combination can ultimately achieve a complex classification of inputs.

The BP neural network is a typical multilayer feedforward neural network. A general BP neural network is divided into two parts:

Forward propagation: In this process, we calculate the final output value and the loss value between the output value and the actual value according to the input sample, given the initialization weight value w, if the loss value is not within the given range The process of backpropagation. If the loss value is within the given range, the update calculation of w is stopped.

Reverse Propagation: Outputs the output of the algorithm form through the hidden layer to the input layer and calculates the error in all cells of each layer. In this way, an error signal for each unit of each layer is obtained, and the error signal is used as a basis for updating the weight value of each unit as a subsequent correction.

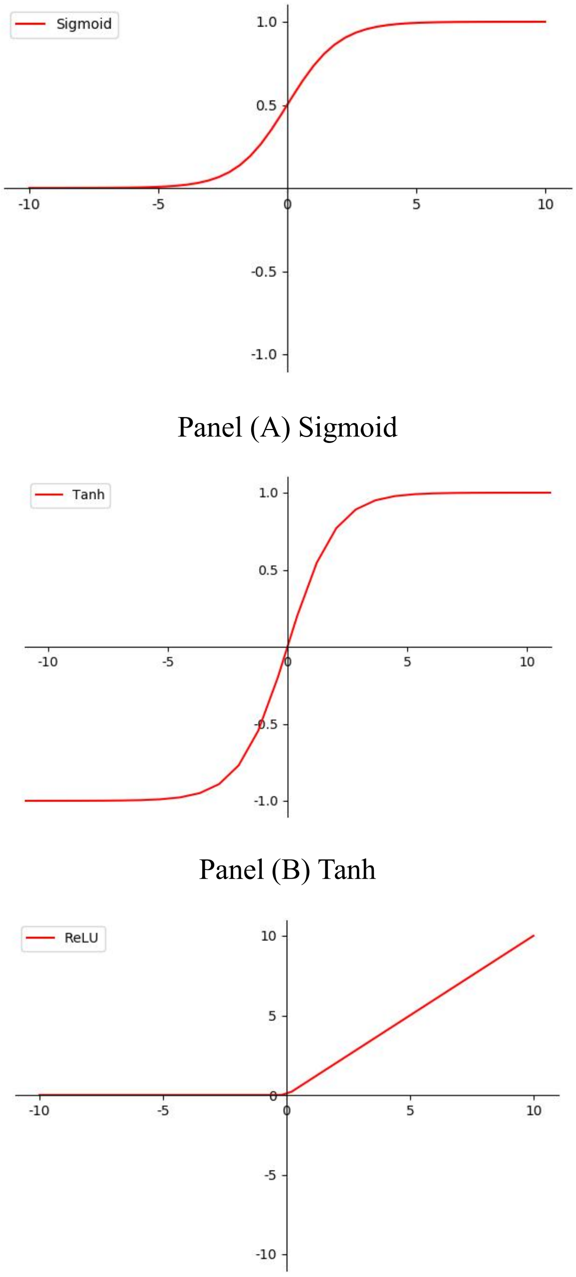
Common activation functions include:

- Sigmoid function

The Sigmoid function is the most commonly used activation function, and its expression is as following equation :

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.31}$$

A schematic diagram of the Sigmod function is shown in Figure 2.1. It converts the continuous real value of the input to an output between 0 and 1. In particular, if it is a very large negative number, the output is 0. If it is a very large positive number, the output is 1. The disadvantage of this function is that gradient inversion and gradient disappearance are caused by gradient back-transfer in deep neural networks. The probability of gradient explosion occurring is very small, and the probability of gradient disappearing is relatively large.



Panel (A) Sigmoid

Panel (B) Tanh

Panel (C) ReLU

Figure 2.1 Activation Function

- Tanh function

The Tanh function is another most commonly used activation function, and its expression is equation :

$$\tan(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.32}$$

A schematic diagram of the Tanh function is shown in Figure 2.1. The disadvantage of this function is that the problem of gradient disappearance still exists.

- ReLU function

The Tan function is also commonly used, and its expression i0s equation :

$$\mathbf{ReLU}(x)=\mathbf{max}(0,x) \tag{2.33}$$

A schematic diagram of the Tan function is shown in following Figure 2.1. Although ReLU is simple, it has the following major advantages: it solves the problem of gradient disappearance, the calculation speed is very fast, and the convergence speed is much faster than the sigmoid and tanh functions.

2.4.6.3 Advantages and disadvantages

The advantages of BP neural network are:

- BP neural network is essentially a mapping from input to output. It has been proved theoretically that the three-layer neural network can approximate any nonlinear continuous function with arbitrary accuracy. This means that BP neural network is especially suitable for solving problems with complex internal mechanism, that is, BP neural network has strong nonlinear mapping ability
- BP neural network has high self-learning and adaptive ability. During training, it automatically extracts the "reasonable rules" between output and output data through learning, and adaptively memorizes the learning content in the weight of the network.
- The network has certain promotion and generalization capabilities.

The disadvantages of BP neural network are:

- The algorithm is slow. There are three main reasons for this problem: the first reason is that the BP algorithm is essentially a gradient descent method, and the objective function to be optimized is very complicated. Therefore, the ‘saw-tooth phenomenon’ is inevitable, which makes the BP algorithm very inefficient. The second reason is that there is paralysis. Because the optimized objective function is very complicated, it will inevitably show some flat areas when the neuron output is close to 0 or 1. In these areas, the weight error changes little, the training process is almost paused. The last reason is that the BP algorithm can't use the traditional one-dimensional search method to find the step size of each iteration, but the step update rule must be given to the network in advance. This method will cause the algorithm to be inefficient.
- Network training is more likely to fail. Mathematically, the BP algorithm is a local search optimization method, but is used to solve the global extremum of the

complex nonlinear function. Therefore, the algorithm is likely to fall into the local extremum and make the training fail. The approximation and promotion ability of BP network is closely related to the typicality of learning samples. It is difficult to select typical sample instances from the problem to form a training set. It is also hard to resolve the conflict between the instance size and the network size of the application problem. This involves the relationship between the possibility and feasibility of network capacity, which isthe problem of learning complexity.

- The most serious problem of BP Neural Network is the inability to explain your reasoning process and reasoning basis. And neural networks typically require more data than traditional machine learning algorithms.

2.4.7 Ensemble learning

Ensemble learning(Polikar, 2006) refers to learning the same problem by different algorithms and integrating various learning results through a certain strategy, in the hope of achieving better learning results than a single algorithm.

The research of ensemble learning is mainly divided into two aspects, one of which is construction of basic classifiers, and the other is the method of integration, such as voting method, stack integration method, cascade combination method and algorithm correlation method. There are two main approaches to construct a single basic classifier.

The first one is Boosting, a method to get higher accuracy though training a rough primary prediction method intensively.

The other is bagging, to acquire training sets through repeated sampling, and enhance individual learning ability and predictive stability. Based on the principal component analysis and SVM advantages, Shen (2011) built a PCA-SVM default integration model for SMEs. The results of the model showed that the hybrid strategy can inherit their respective advantages, and can discriminate precisely with only a few indicators, which provides new ideas and empirical results for commercial banks to study quantitative models. L. M. Wang et al. (2016) proposed to repeatedly sample five–seed-models using decision tree method, BP neural network, SVM, linear programming, Naive Bayes classification method, and continuously revised the weight of each model according to the obtained accuracy, and built the credit evaluation model of SMEs based on the ensemble learning theory. The empirical results showed that the model can avoid the problem of over-fitting of a single model, and has stronger generalization ability and higher prediction accuracy.

It is found that only when the results of the various algorithms involved in the integration were divergent, the error rate of each algorithm was less than 0.5, and the errors were mutually independent, the ensemble learning based on the various algorithms could complement each other and improve the operational performance and classification effect.

Ensemble learning is a very important algorithmic type. Then we will introduce three important ensemble learning method: random forest,ADABOOST and XHBOOST.

2.4.8 Random forest

2.4.8.1 Brief introduction

The high cost of computation to prune a Decision Tree always causes purity and accuracy problems. In practical application, Random Forest algorithm can construct a Decision Tree Forest with voting function to deal with the Decision tree's problem of over-fitting, by randomly choose several small Decision Trees to combine with, which needs only a small part of computation. In the process of making classification decisions, each Decision Tree of the forest will have a prediction on the classification of the sample set, and the voting function will vote for each result and conclude the final results of classification(Ye&Lu, 2017).

Suppose that there are m Decision Trees, each has a result of classification c_i, then the final result will be equation :

$$\boldsymbol{I} = \mathbf{arg\,max}\left(\boldsymbol{p_i}\right) \; \boldsymbol{where\, p_i} = \frac{\boldsymbol{c_i}}{\boldsymbol{m}} \tag{2.34}$$

We need to choose m sample sets with n samples as the entire sample set to train the Random Forest model. The reason of not using the whole sample as the training set is mainly because the cost of computation is too high and some certain characteristics of feature attributes in local sample sets will be ignored in that case, which will cause drops in generalization ability.

Therefore, Random Forest algorithm uses a method of random sampling, randomly choosing subsets from the original sample set and conducting trainings on each of them with the Decision Tree model. The algorithm can not only avoid the problem of over-fitting, but also compare different Decision Trees trained by different sample sets and find feature attributes that are important to classification.

2.4.8.2 Construction of a random forest

Randomly choose samples with replacement. First, determine the number of samples in the training set according to the capacity of the entire sample set. Then using the strategy of random sampling with replacement, keep choosing several training sets, aware that samples in different training sets can be the same since we are doing the sampling with replacement.

Randomly choose feature attributes for splitting without replacement. When constructing the CART Decision Tree, we only need to choose feature attributes to split the sample sets from the entire sample set, calculate the Gini coefficient of the chosen attribute and then get the optimal feature attribute for splitting.

Aware that we are sampling feature attributes without replacement, so there is no need to worry about duplications in the chosen attributes. Random Forest algorithm contains two random processes: randomly sampling sample subsets with replacement and randomly sampling feature attributes without replacement.

Vote for each Decision Tree. By executing the two processes mentioned above, one of the Decision Trees in the forest can be constructed and by repeating these processes, several Decision Trees will combine and finally a Random Forest will be obtained. When classifying one of the samples in the sample set, Random Forest algorithm will get several results of classification of the sample by Decision Trees. Then we sort all the results according to the votes they get in the forest and finally conclude the prediction by Random Forest algorithm.

2.4.8.3 Advantages and disadvantages

Random Forest algorithm adds the process of random sampling without losing the advantages of the Decision Tree model and form a voting pool by combining results of several Decision Trees.

It has several advantages as follows:

- It has a higher accuracy of classification in practical application.
- The Random Forest algorithm choose feature attributes by random sampling so it doesn’t need any feature project to screen the feature attributes, while it can deal with a large scale of complex data with high dimensions.
- After random forest training, we can clarify the importance of each feature attribute. Although the characteristic attributes used by each decision tree are different, the results of the decision tree can be evaluated by final classification and voting. The closer the result of the decision tree is to the final result, the better the training of the model.
- The estimated value of generalized error in Random Forest algorithm is unbiased. That is to say it has a strong generalization ability which is unlikely to cause over-fitting problem.

- The cost of training a Random Forest is relatively low. Under the help of distributed computing, it can get the final model rapidly, which is a great advance compared withDecision Tree models.
- Random Forest models are not sensitive to missing values. Even if values of some parts of samples' feature attributes are missing, the algorithm can still maintain its accuracy. This is because the process of choosing feature attributes for splitting is equal to a process of removing some attributes. Each Decision Tree only has a small part of attributes of the entire sample set, which then means parts of missing won't affect the accuracy of final results.

However, as the essence of a Random Forest is a combination of Decision Trees, it still has the inherent problems of Decision Tree models.

Random Forest algorithm is appropriate to sample sets with a large scale of samples and variables. It's more likely to have over-fitting problems when dealing with a sample set with relatively less feature attributes. The ability of recognizing noisy of Random Forest algorithm is relatively weaker than other machine learning algorithms(J. G. Gao et al., 2015).

2.4.9 XGBOOST

2.4.9.1 Brief Introduction

XGBOOST is a boosting tree algorithm proposed byT. Chen &Guestrin(2016). It can perform multi-thread parallel computing. It generates generations of new trees through iterations.

It actually combines many weak learners with low classification performance into one accurate one. A strong learner with a high rate, each decision tree may not have a good classification effect, but the results of multiple classifications will be more accurately predicted. In order to find the optimal solution, XGBOOST adds regular items to the objective function, which reduces the complexity of the objective function and the model, avoids over-fitting, and has the advantages of fast running speed, good classification effect, and support for custom loss function.

2.4.9.2 Algorithm

The basis of the XGBOOST algorithm is the GBDT algorithm. The goal of each round of optimization in GBDT is to calculate the loss function of the previous tree output values and use the loss function to fit. The general loss function of GBDT is in the form of a log-likelihood function, which is equation:

$$L\left(y,f\left(x\right)\right)=\log\left(1+\exp\left(-yf\left(x\right)\right)\right);y=\left\{-1,1\right\} \tag{2.35}$$

Calculate the negative gradient,which is equation :

$$r_{t,i}=-\left[\frac{\partial L\left(y_i,f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f\left(x_i\right)=f_{t-1}\left(x_i\right)}=\frac{y_i}{1+\exp(y_i\left(f_{t-1}\left(x_i\right)\right)} \tag{2.36}$$

Using a gradient to fit a CART regression tree, the t-th regression tree is obtained, and the corresponding leaf node region is $R_{t,j}$, $j=\{1,2,...,J\}$. where J is the number of leaf nodes of the regression tree t. Then we need to estimate the value at each leaf node, ie the best residual fit value for each leaf node $c_{t,i}$,as equation shows:

$$c_{t,j}=\arg\min\sum\log\left(1+\exp\left(-y_i\left(f_{t-1}\left(x_i\right)+c_{t,i}\right)\right)\right) \tag{2.37}$$

This formula is difficult to optimize, so the gradient is approximated by the equation :

$$c_{t,j}=\frac{\sum_{xi\in R_{t,j}}r_{t,i}}{\sum_{xi\in R_{t,j}}\left|r_{t,i}\right|\left(1-\left|r_{t,i}\right|\right)} \tag{2.38}$$

After obtaining$c_{t,i}$, the learner can be updated by using the rule of adding residuals, which is expressed as equation :

$$f_t\left(x_i\right)=f_{t-1}\left(x_i\right)+\sum_{j=1}^{J}c_{t,j}I_{t,i}\left(x_i\in R_{t,j}\right) \tag{2.39}$$

Finally we can get a strong learner which is equation :

$$f\left(x\right)=f_T\left(x\right)=f_0\left(x\right)+\sum_{t=1}^{T}\sum_{j=1}^{J}c_{t,j}I_{t,i}\left(xi\in R_{t,j}\right) \tag{2.40}$$

For XGBOOST, the loss function has one more penalty than the GBDT loss function, defined as equation :

$$\begin{gathered}L(\varnothing)=\sum_{i=1}^{n}l\left(\hat{y}_i,y_i\right)+\sum_{k=1}^{K}\Omega\left(f_k\right)\\ \Omega\left(f_k\right)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2\end{gathered} \tag{2.41}$$

Where $\hat{y}_i$ is the predicted value, $\Omega(f_k)$ represents the model complexity metric function, *k* is the number of trees, f_k is the kth tree, *n* is the number of samples, w_j is the leaf node score of the tree, T is The number of leaf nodes. The objective function requires that the residuals be as small as possible and the number of leaf nodes as small as possible. The parameters of the tree are obtained through a greedy strategy (determined by the immediate benefit maximization), that is, for each classification, only the benefit of the current classification is maximized.

The following is a derivation of how to split the tree. Unlike the GBDT algorithm, XGBOOT does not use the mean, but obtains the leaf node score through the second derivative. The first is the score of the leaf node. for the next tree, the optimization goal is equation :

$$\begin{aligned} Obj &= \sum_{i=1}^{n} l(\hat{y}_{i,t}, y_{i,t}) + \sum_{k=1}^{K} \Omega(f_k) = \sum_{i=1}^{n} l\left(\hat{y}_{i,t-1} + f_t\left(x_i\right), y_i\right) + \Omega\left(f_k\right) + C \\ &= \sum_{i=1}^{n}\left(f_t\left(x_i\right)^2 + 2\left(\hat{y}_{i,t-1} - y_i\right) f_t\left(x_i\right)\right) + \left(\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2\right) + C1 \\ &\approx \sum_{i=1}^{n}\left[l\left(\hat{y}_{i,t-1}, y_{i,t}\right) + \frac{\partial l\left(\hat{y}_{i,t-1}, y_{i,t}\right)}{\partial \hat{y}_{i,t-1}} f_t\left(x_i\right) + \frac{\partial^2 l\left(\hat{y}_{i,t-1}, y_{i,t}\right)}{2\partial \hat{y}_{i,t-1}{}^2} f_t\left(x_i\right)^2 \right] + \Omega\left(f_k\right) + C \\ &= \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t\left(x_i\right)^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + C1 \\ &= \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j{}^2 \right] + \gamma T \end{aligned} \tag{2.42}$$

Find the partial derivative of w_j to make it 0, which is equation :

$$\begin{aligned} G_j w_j + \left(H_j + \lambda\right) w_j &= 0 \\ w_j^* &= -\frac{G_j}{H_j + \lambda} \\ Obj^* &= -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \gamma T \end{aligned} \tag{2.43}$$

The following is a derivation of how to split. For any split, the gain is expressed as equation :

$$\begin{aligned} Gain &= \frac{1}{2}\left[\frac{G_L{}^2}{H_L + \lambda} + \frac{G_R{}^2}{H_R + \lambda} - \frac{\left(G_L + G_R\right)^2}{H_L + H_R + \lambda} \right] - \gamma \\ &= \frac{1}{2}\left[\textit{left tree loss} + \textit{right tree loss} - \textit{undivided loss} \right] - \textit{new node added penalty} \end{aligned} \tag{2.44}$$

2.4.9.3 Advantages and disadvantages

The XGBOOST method has the following advantages:

- Use of the Taylor expansion method (considering first-order and second-order derivatives) for approximate estimation when performing node scores, and GBDT is the first-order derivative used.
- XGBOOST considers the case where the training data is sparse. A variety of methods to prevent overfitting are used: pruning, sample sampling, feature sampling, Shrinkage.
- XGBOOST adds the processing of missing values. XGBOOST can automatically

learn the splitting direction.

2.4.10ADABOOST

2.4.10.1 Brief introduction

Boosting is similar to the Bagging method in that the combined output of multiple models is realized either by voting (for classification) or by averaging (for numerical prediction). Another similarity is that the Boosting algorithm also requires that the combined classification model can only be one model, not multiple models. The difference with Bagging method is that the construction process of each classification model is circular and iterative, while Boosting method is independent in the establishment of each classification model.

Therefore, Boosting method each generated classification model is affected by the performance of the last established classification model. Boosting's approach weights the contributions of each classification model according to the performance of the classification model, where the weight of each model is the same in the traditional bagging algorithm(Hastie, et al, 2009; Rosset, et al, 2009; D. Li et al., 2016).

Boosting method is one of the most effective learning algorithms in the last decade, with the idea of merging the outputs of multiple weak learning models to generate a "committee". This process is somewhat similar to the Bagging method, but in essence they are different

ADABOOST (Hastie et al., 2009) is a typical Boosting algorithm. The Boosting algorithm is a process of upgrading the ‘weak learning algorithm’ to the ‘strong learning algorithm’. The main idea is that ‘three smugglers are the top ones.’

The Boosting algorithm involves two parts, an addition model and a forward step-by-step algorithm. The addition model means that the strong classifier is linearly added by a series of weak classifiers. TheADABOOST weighting makes the weak classifier with high correct rate and good partitioning effect more weight. Forward step-by-step means that during the training process, the classifier generated in the next iteration is trained on the basis of the previous round.

ADABOOST changes the weight of the training sample in the loss function, and its idea is to pay attention to the misclassified sample in the classifier, it increases the weight of the misjudged sample and reduces the weight of the correct sample, thus ensuring that the learning level of the wrong sample is continuously improved during the training process, and the guiding model continuously corrects its error. Thereby obtaining more accurate prediction results.

2.4.10.2 Algorithm

The residual learning process ofADABOOST is more complicated. In simple terms, theADABOOST weighting makes the weight of the weak classifier with high correct rate and good partitioning effect larger, at the same time,ADABOOST changes the weight of training data, the idea is to put the focus on a sample that was misclassified in a round of classifiers.

- Input: $\boldsymbol{T} = \{(\boldsymbol{x_1}, \boldsymbol{y_1}), (\boldsymbol{x_2}, \boldsymbol{y_2})......(\boldsymbol{x_N}, \boldsymbol{y_N})\}$, $\boldsymbol{y_i} = \{\mathbf{1}, -\mathbf{1}\}$
- Initialize the original weight: $\boldsymbol{D_1} = (\boldsymbol{w_{1,1}}, \boldsymbol{w_{1,2}}, ..., \boldsymbol{w_{1,N}})$, $\boldsymbol{w_{1,1}} = \frac{\mathbf{1}}{\boldsymbol{N}}$ wherem represents iterations, and $\boldsymbol{m} = \mathbf{1}, \mathbf{2}, ..., \boldsymbol{M}$
- Learning with the training dataset of the weight distribution D_m, the weak classifier $\boldsymbol{G_m}(\boldsymbol{x})$ is obtained.
  a) Calculate the classification error rate of $\boldsymbol{G_m}(\boldsymbol{x})$:$\boldsymbol{e_m} = \sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{w_{m,i}} \boldsymbol{I}(\boldsymbol{G_m}(\boldsymbol{x_i}) \neq \boldsymbol{y_i})$
  b) Calculate the weight of the $\boldsymbol{G_m}(\boldsymbol{x})$ classifier:$\boldsymbol{a_m} = \frac{\mathbf{1}}{\mathbf{2}} \mathbf{log}\,(\frac{1-e_m}{e_m})$
  c) Update the weight distribution of the training dataset: $\boldsymbol{w_{m+1,i}} = \frac{\boldsymbol{w_{m,i}}}{\boldsymbol{Z_m}} \mathbf{exp}\,(-\,\boldsymbol{a_m y_i G_m}(\boldsymbol{x_i}))$, $\boldsymbol{Z_m}$ is the normalization factor
- Finally we can get the classifier:$\mathbf{F(x)} = \mathbf{sign}(\sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{a_m G_m}(\boldsymbol{x}))$

It can be seen that the weight update of the ADABOOST includes the weak classifier weight a_m and the sample weight $w_{m,i}$. The following is the first to derive the weak classifier weight update formula:

For *m* round iterations, $\mathbf{F}_{\boldsymbol{m}}(\boldsymbol{x}) = \mathbf{F}_{\boldsymbol{m}-\mathbf{1}}(\boldsymbol{x}) + \boldsymbol{a_m G_m}(\boldsymbol{x})$, the purpose of the algorithm is to minimize $\mathbf{F}_{\boldsymbol{m}}(\boldsymbol{x})$. Minimizing the language expression of the algorithm is to minimize the loss function. The loss function is equation :

$$\begin{aligned} \boldsymbol{Loss} &= \sum_{i=1}^{N} \exp\left[-y_i\left(F_{m-1}(x) + a_m G_m(x)\right)\right] = \sum_{i=1}^{N} w_{m,i} \exp\left(-y_i a_m G_m(x_i)\right) \\ &= \sum_{y_i = G_m(x_i)}^{N} w_{m,i} \exp(-a_m) + \sum_{y_i \neq G_m(x_i)}^{N} w_{m,i} \exp(a_m) \\ &= \sum_{i=1}^{N} w_{m,i} \left(\frac{\sum_{y_i = G_m(x_i)}^{N} w_{m,i}}{\sum_{i=1}^{N} w_{m,i}} \exp(-a_m) + \frac{\sum_{y_i \neq G_m(x_i)}^{N} w_{m,i}}{\sum_{i=1}^{N} w_{m,i}} \exp(a_m) \right) \\ &= \sum_{i=1}^{N} (1 - e_m) \exp(-a_m) + e_m \exp(a_m) \end{aligned} \tag{2.45}$$

$$w_{m,i} = \exp\left(-y_i F_{m-1}(x)\right) = w_{m-1,i} \exp\left(-y_i a_{m-1} G_{m-1}(x_i)\right)$$

Loss seeks partial bias on a_m to make it 0, which gives equation :

$$a_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \tag{2.46}$$

Then derive the sample weight update formula,the above is assuming that the weight update of $w_{m,i}$ is known, and the optimization of a_m is derived. Now it is necessary to assume that a_m is known, and the optimization of $w_{m,i}$ is derived. Therefore, the Loss function is expressed in the desired form, and the corresponding distribution is expected to be the weight of the sample (with Taylor expansion),which can be written as equation :

$$\begin{aligned} Loss &= \sum_{i=1}^{N} \exp\left[-y_i(F_{m-1}(x)+a_m G_m(x))\right] = E\left[\exp\left(-y_i(F_{m-1}(x)+G_m(x))\right)\right] \\ &\approx E\left[exp\left(-y_i(F_{m-1}(x)\right)*\left(1-y_iG_m(x)+\frac{1}{2}y_i^2G_m^2(x)\right)\right] \\ &= E\left[exp\left(-y_i(F_{m-1}(x)\right)*\left(1-y_iG_m(x)+\frac{1}{2}\right)\right] \end{aligned} \tag{2.46}$$

For an ideal learning period, you need to minimize Loss, which is consistent with the above ideas and written as equation :

$$\begin{aligned} \min Loss &= \max E\left[exp\left(-y_i(F_{m-1}(x)\right)\times y_iG_m(x)\right] \\ &= \max E\left[\frac{exp\left(-y_i(F_{m-1}(x)\right)}{E\left(exp\left(-y_i(F_{m-1}(x)\right)\right)}\times y_iG_m(x)\right] \\ &= \max E_{D_t}\left[y_iG_m(x)\right] \end{aligned} \tag{2.47}$$

The original distribution is D(x), that is, the sample weight that is expected to be used is $w_{1,i}$, the sample weight after the change is written as equation :

$$\begin{aligned} D_t(\mathrm{x}) &= \frac{D(x)exp\left(-y_i(F_{m-1}(x)\right)}{E\left(exp\left(-y_i(F_{m-1}(x)\right)\right)} \\ y_iG_m(x) &= 1-2I\left(y_i \neq G_m(x)\right) \\ \min Loss &= \min E\left[y_i \neq G_m(x)\right] \end{aligned} \tag{2.48}$$

It can be seen that minimizing the classification error under the distribution of $D_t(\mathrm{x})$ is equivalent to learning the ideal classifier, or training by $D_t(\mathrm{x})$, and the ideal classifier can be obtained from equation :

$$\begin{aligned} D_{t+1}(\mathrm{x}) &= \frac{D(x)\exp\left(-y_i(F_m(x)\right)}{E\left(\exp\left(-y_i(F_m(x)\right)\right)} = \frac{D(x)\exp\left(-y_i\left(F_{m-1}(x)+a_mG_m(x)\right)\right)}{E\left(\exp\left(-y_i(F_m(x)\right)\right)} \\ &= D_t(\mathrm{x})\exp\left(-y_ia_mG_m(x)\right)\frac{E\left(\exp\left(-y_iF_{m-1}(x)\right)\right)}{E\left(\exp\left(-y_i(F_m(x)\right)\right)} \\ &= D_t(\mathrm{x})\exp\left(-y_ia_mG_m(x)\right)\frac{1}{E\left(\exp\left(-y_ia_mG_m(x_i)\right)\right)} \end{aligned} \tag{2.49}$$

2.4.10.3 Advantages and disadvantages

First of all,ADABOOST could lead to over-fitting, but it is not particularly serious. Here, the

SVM method is taken as an example for comparison. In practice, the generalization performance ofADABOOST is generally inferior to that of SVM. Therefore,ADABOOST is preferred when only a large number of features need to be processed or feature selection is required, and the speed is required. Linear SVMs have much higher generalization performance but are very inefficient in the case of high dimensional data. Secondly,ADABOOST is to increase the weak classifier, the upper bound of the training error will continue to decline, and the final model will be obtained. Therefore, the update of the weight is very sensitive to the accuracy of the sample label. Individual error labels will greatly affect the overall model performance.

2.4.11 Summary

This section reviews the similarities and differences, advantages and disadvantages of many machine learning models, and discusses their application in the construction of enterprise credit evaluation models.

Logistic regression is the simplest and widely used linear method. It is a method to analyze enterprises' credit status based on discrete data, the prediction accuracy of which is commonly thought to be between 54% and 90% at present(Peng, 2012; Qing &Xin, 2015). However, this method has a lot of shortcomings: First, Logistic regression is only useful for predicting discrete functions and cannot predict continuous functions; Second, Logistic regression requires high data quality and sample size, which is not only sensitive to outliers, but also requires that there is no collinearity between independent variables; Third, Logistic regression can only represent the "point-to-point" logical relationship, which is difficult to apply to complex logical relationships, so it is limited compared with other complex methods such as neural networks (Singh et al., 2021)。

The application of decision tree theory to credit evaluation of small and micro enterprises is also an early stage of research. Compared with Logistic regression, it has the advantage that continuous variables can be classified and predicted, and there is no need to make any specific assumptions about the distribution of variables. Therefore, when there are many missing data and dimensional differences, the decision tree theory is more applicable, but its disadvantage is that it is extremely prone to over fitting problems, Therefore, this method is rarely used at this stage (Satchidananda et al., 2006; Bhattacharya et al., 2022)。

Compared with Logistic regression and decision tree theory, naive bayes model has a solid mathematical foundation and a high but stable classification efficiency, which not only overcomes the shortcoming of Logistic regression that is not suitable for small samples, but also overcomes the shortcoming of decision tree that cannot handle multiple classification tasks, and is suitable for incremental training. At the same time, the algorithm of naive bayes model is simple and does not require high data. However, its disadvantage is that it needs to calculate a priori probability, and there is a certain error rate. Therefore, the accuracy of credit evaluation using naive Bayesian method is low (Aithal et al., 2019)。

Integrated learning is not a single machine learning algorithm, but a learning task completed by building and combining multiple machine learners, so it can be regarded as a meta algorithm. In terms of specific composition, for training set data, a strong learner can be finally formed by training several individual weak learners and certain combination strategies. In recent years, integrated learning has gradually become a hot research field. It can improve the generalization ability of the algorithm by repeatedly sampling or integrating weak learning, and has good practical effect. According to the working mechanism, it can be divided into Bagging, Stacking and Boosting.。

Among them, Bagging conducts sub sampling from the training set to form the sub training set required by each base model, and then averages the predicted results of all base models to produce the final prediction results. The core of Bagging is how to sample. Generally, self-service method is used for sampling. The representative of this mechanism is random forest, which is equivalent to higher level Bagging. Because the weak learners of random forests are decision trees, the sample features are randomly selected based on the self-service sampling of Bagging samples. The advantage of random forest is that it can be used in a wider range, and as long as it contains enough decision trees, it is not easy to be affected by the over fitting problem. However, the training cost of random forests is positively related to the number of decision trees, which requires users to balance the training cost and training accuracy. In addition, when the sample noise is too large, it is easy to have the problem of over fitting. But on the whole, random forest is more effective than Logistic regression and decision tree in evaluating enterprise credit(Hamori et al., 2018; Uddin et al., 2022)。

In contrast, Stacking is to stack various classifiers, that is to say, after training all the base models, it is necessary to use the training results to predict the training base, and meet the symmetry. The prediction value of the i base model for the j training sample will be used as the i characteristic value of the j sample in the new training set, so as to form a new training set and train on this basis. Similarly, the prediction process is to form a new test set through the prediction of all the base models, and finally drop the new test set for prediction. This method is similar to the upgrading of Bagging, including KNN algorithm and SVM. KNN method is based on a simple idea: the characteristics of samples of the same type are often similar. Therefore, there is no cost calculation in the learning process. All calculations are completed at the time of prediction, and no assumptions are made about the data distribution. However, KNN algorithm has obvious shortcomings: First, this method can not directly deal with classification variables, and can not be used for high-dimensional data sets; Secondly, it is prone to dimensional disasters. When the training set is too large, the workload of calculating the distance between new data and all samples in the training set will increase exponentially; Third, the prediction accuracy is sensitive to noise and abnormal values。Therefore, KNN is often only used for simple classification and short-term credit risk assessment (Abdelmoula, 2015).In contrast, SVM is suitable for evaluating enterprise credit in high-dimensional, nonlinear and small sample conditions, and the more characteristic variables, the higher the accuracy. However, its disadvantage is that the calculation cost is high, and multiple super parameters need to be adjusted at the same time during calculation, which is sensitive to parameter selection, and generally can only be used to deal with continuous variables (Yu et.al, 2010).

ADABOOST and XGBOOST belong to Boosting algorithm in integrated learning, while random forest belongs to Bagging algorithm in integrated learning. The main advantage of these three algorithms is that they can better fit samples and use multiple weak classifiers. However, different weak classifiers in a random forest are equal. In ADABOOST and XGBOOST algorithms, there is a progressive relationship between consecutive weak classifiers. The latter weak classifier partially depends on the classification effect of the previous weak classifier. In practical application, the random forest algorithm can only be used in classification scenarios. ADABOOST and XGBOOST algorithms can not only be used for classification, but also for predicting the specific value of the target(Gao et.al, 2010).

Neural network and integrated learning intersect each other. In integrated learning, a basic learner in the form of neural network can be used. The multi-channel in neural network method is also similar to the integrated idea of integrated learning. BP neural network model is a very typical nonlinear neural modeling method. Because of its high requirements for the computing ability of computers, it was not used much in the early years. But in recent years, due to the progress of computer technology, it has received more and more attention in theory and application. In fact, BP neural network model is prone to over fitting when there are too many parameters, so it is mainly used for sequence prediction rather than classification prediction (Huang et.al, 2010).

2.5 Feature selection

The effect of the credit scoring model depends not only on the design of the model, but also on the selection of indicator variables in the model. For SMEs, information acquisition and credit risk assessment are difficult due to their historical archives are not perfect and information transparency is low(B. L. Fan & Zhu, 2003).

Most studies only consider the financial situation of the company, which is not enough(Cooper et al., 1991;P. Liu & Shen, 2012). First of all, the accounting system of SMEs is not perfect, and the credibility of its financial data is not high. Secondly, for most SMEs in China, the role of qualitative indicators of enterprises and their responsible persons cannot be ignored(Niu, 2005).

Therefore, it is necessary to consider not only the financial situation of the enterprise, but also other information of the enterprise, such as the business owner (H. Zhang 2008). In addition, indicators that reflect industry risks and the business environment are also important. That is, the characteristic variables of enterprises needto include the impact of macro-economy on enterprises, the size of enterprises and the characteristics of their industries (Qiu&Chen, 2014; Zheng, 2015).

All enterprises are in a certain macroeconomic environment. The overall economic development speed and stability of a country or region will have varying degrees of impact on all walks of life. Therefore, we will assess the changes in the macroeconomic environment. Consider the impact of target product or service demand, raw material supply, profitability and asset quality on enterprise credit

If the enterprise is small, its operation and development are more vulnerable to the whole industry environment, regional economic and institutional environment (P. Liu&Shen, 2012; Wu, 2013). The regional economic system environment faced by enterprises, such as economic development level, macroeconomic prosperity, market-oriented development, public goods supply, local supervision and industrial accumulation, will affect the operation ability and solvency of enterprises.

The information transparency of large enterprises is high, and the unit cost of bank supervision is relatively small. Therefore, the larger the scale of the enterprise, the smaller the possibility of its credit default. The nature of the industry in which the enterprise is located will also affect the possibility of credit default. Different industries have different profitability and risk, and agriculture is weak. Therefore, agricultural enterprises are more likely to default than non-agricultural enterprises(Qiu & Chen, 2014).

The goal and dataset in the historical literature are summarized in Table 2.2Panel (A)and non-financial indicators are summarized in Table 2.2Panel (B). The reason for not paying attention to financial indicators is that information acquisition and credit risk assessment are difficult due to their historical archives are not perfect and information transparency is low for SMEs. More seriously, SMEs do not have a sound financial system and audit environment, so the financial data are unreliable. So non-financial indicators are more important for SMEs' credit evolution.

Table 2.2 List of related literatureand summary of variables

Panel (A)

Reference	**No**	**Goal**	**Dataset**	**Num of Sample**	**Num of Feature**
L. J. Gao 2012	A1	Evaluate company's solvency	Credit Reform 1996-2004	Unknown	16
Wu 2013	A2	The impact of growth factors on solvency	A-share companies 2007-2009	100	9
Zhou & Wang 2014	A3	Effectiveness of risk factors	SMEs Lending Data of Beijing 2004-2007	193	13
P.Liu et al. 2012	A4	Design a credit evaluation scoring table	Unknown	Unknown	24
Huo 2012	A5	Credit risk of high-tech SMEs	Unknown	Unknown	19
Cooper et al. 1991	A6	New venture survival.	Surveys, Short-response postcards, post-office return mail, NFIB checks	2994	11
Yeh et al. 2012	A7	Present a credit rating prediction model	Taiwan Economic Journal (Securities Markets)	2570	21
Steijvers et al. 2005	A9	A Decision Tree Analysis of	NSSBF (Companies fewer than 500 employees)	Unknown	6

Panel (B)

Variables	A1	A2	A3	A4	A5	A6	A7	A8	A9
Loan availability									
Assets		1		1		1		1	1
Staff size									
Industry Development Prospect								1	
Macroeconomic								1	
Location									
Enterprise type						1			
Scope of business				1				1	
Solvency	1								
Num of years of business existence	1								
Num of days before reimbursement	1								
Number of employees	1		1						
Payment history	1			1				1	
Margin profit			1			1			
Enterprise age			1						

Years of cooperation with Banks	1						1
Quality of management personnel		1	1	1			
Quality of employees		1	1			1	
R&D spending			1				
Number of R&D staff			1				
Manager gender				1			
Whether the administrator is an ethnic minority				1			
Number of Full-time Partners				1			
External guarantee						1	
The number of banks the firm negotiates with before agreeing to a certain credit contract							1
Family business							1
Loan amount							1

Chapter 3: Methodology

This chapter introduces the research methods, the basic information of the dataset and the problems to be solved. Among them, Section 3.1 will introduce the dataset and sample segmentation method; in Section 3.2, the main machine learning methods used in this study will be presented; in Section 3.3, a general solution to the imbalance problem in the dataset after analysis will be given. In Section 3.4, the potential impact of another important issue: the fact that the labels used may not be trusted will be highlighted

3.1 Data and sample

3.1.1 Dataset

This study uses the record data of loan details of SMEs from between 2011 and 2016 from a subbranch of a bank, with a total number of 848 samples. By eliminating records which use USD as loan currency, 846 samples remained, which is so small to train complex deep learning model. As this data set is the real data collected by commercial bank, it will be opened to the academic community after technical processing to hide enterprise privacy for the use of small and micro enterprises.That is also the reason for just testing more general and simple ML models in our study. Then ‘Positive’ labels are added to normal loans and ‘Negative’ labels to bad loans, according to their field ‘Risk Classification Result’ in the loan record. Finally, 78 negative samples and 768 positive samples among the original samples are obtained, which means a ratio of 9.846:1 between positive and negative samples.

Due to the serious imbalance of data, the loss function will be biased to the side with more samples during training, resulting in small loss function value and low recognition accuracy for categories with small sample size. This problem can be solved by expanding a few sample categories or adjusting the loss function. This study uses the former method because the former can be adjusted at one time without adjusting the loss function of each model separately.

After enlarging minority category, there are two operations for independent variables: Firstly, transform multi-categories variable into multiple dummy variables. Secondly, standardize continuous variables with maximum value 1 and minimum value 0. Table 3.1

shows the variable type and the variable mapping.

Table 3.1 Variable description

Panel (A)

Variable	Meaning	Variable types
Y	Label (Negative/Positive)	Binary dummy
DKZL	Type of Loan	Multiply dummy
QYGM	Company Size	Multiply dummy
HKFS	Method of Repayment	Multiply dummy
YWPZ	Business Varieties	Multiply dummy
XXDBFS	Detailed Method of Guarantee	Multiply dummy
FFJE	Amount of Loan	Continuous variable
BJJE	Remaining Unpaid Principal	Continuous variable
LLFDZ	Floating Interest Rate	Continuous variable
ZXNLL	Real Interest Rate	Continuous variable
Maturity	Time Until Maturity Date	Continuous variable
Life	Life of the Loan	Continuous variable
Registered Capital	Registered Capital	Continuous variable
Company Industry	Industry	Multiply dummy

Panel (B)

	Total	Mean	Std	Min	50%	Max
FFJE	1542	15483375	43154507	28000	5000000	5E+08
BJJE	1542	13674570	39171814	28000	5000000	5E+08
LLFDZ	1542	45.02377	31.43031	-30	35	244.83
ZXNLL	1542	8.437497	1.380739	3.92	8.53	15
Maturity	1542	43.29053	531.4875	-1707	140	3506
Life	1542	522.6096	452.5773	0	344	2071
Registered Capital	1542	43631984	2.17E+08	75000	6510000	1.84E+09

Panel (C)

	Total	Num	Min	50%	Max
DKZL	1542	17	2	14.5	694
QYGM	1542	4	27	206	805
HKFS	1542	4	4	121	126
YWPZ	1542	16	1	20	185
XXDBFS	1542	17	1	19.5	421
Company Industry	1542	14	1	17	594

Panel (D)

Variables	A1	A2	A3	A4	A5	A6	A7	A8	A9	This study
Assets size		1		1		1		1	1	1
Staff size										
Industry Development Prospect								1		1
Macroeconomic								1		
Location										
Enterprise type						1				
Scope of business				1				1		1
Solvency	1									
Num of days before reimbursement	1									1

Variables	A1	A2	A3	A4	A5	A6	A7	A8	A9	This study
Number of employees	1		1							
Payment history	1			1				1		
Margin profit			1			1				
Enterprise age	1		1							
Years of cooperation with Banks			1						1	
Quality of management personnel				1	1	1				
Quality of employees				1	1			1		
R&D spending					1					
Number of R&D staff					1					
Manager gender						1				
Whether the administrator is an ethnic minority						1				
Number of Full-time Partners						1				
External guarantee								1		1
The number of banks the firm negotiates with before agreeing to a certain credit contract									1	
Family business									1	
Loan amount									1	1
Interest Rate										1
Loan type										1
Method of Repayment										1
Remaining Unpaid Principal										1
Registered Capital										1
Business Varieties										1

Table 3.1 Panel (C) show the statistics information about dummy variables. Num represents the number of variable categories. Rank number of samples for all categories of a variables from smallest to biggest, and Min, 50%, Max represent the 0, 50%, 100% quantile of these samples size.

3.1.2 Variable comparison

Table 3.1 Panel (D) compares variables in this study with other literature. The current literature does not agree on which variables should be used in the credit evaluation model. The variables selected in different literatures are very different.

The main reason for this phenomenon is that data for many variables are not available in SMEs. Some enterprises have data for some special variables but other haven't or they are unreliable. Only Assets size, Scope of business, Payment history, Quality of management personnel and Quality of employees are used in the literature more than twice. Only Assets size is agreed for most of literature. Information like Quality of management personnel and Quality of employees is special and rare, which may be useful and have more explanatory power, but only can be used for rare enterprises.

The advantage of the dataset used in this study is that the information which must be registered for Chinese banks is relatively reliable. Some variables which were used in past research are not presented due to the fact that they are not universal variables and usually appear data missing.

3.1.3 Sample splitting

An important preliminary step (before discussing specific models and regularization methods) is to understand how we design disjoint subsamples for estimation and testing, and to introduce the concept of hyperparameter tuning. The regularization program discussed below is the major measure of machine learning to prevent overfitting, and it relies on the choice of hyperparameters (or tuning parameters).

This is critical to the performance of machine learning methods because it determines the complexity of the model. Hyperparameters include: penalty parameters in lasso and elastic net, the number of iterative trees in boosting, the number of random trees in the forest, and the depth of the tree (Gu et.al, 2020).

In most cases, for how to tune hyperparameters to optimize out-of-sample performance, there are few theories that guide us to follow the most common methods in the literature and adaptively select tuning parameters from the data in the validation samples. In particular, we divide the sample into three non-intersecting sub-samples according to time and maintain the time series of the data. The first sub-sample (or training sub-sample) is used to estimate the model based on a set of specific tuning parameter values.

The second sub-sample (validation) is used to tune hyperparameters. We predict the data points in the validation sample based on the estimation model of the training sample. Next, we calculate the objective function based on the prediction error of the verification sample, and iteratively search and optimize the hyperparameters of the verification target (each step re-estimates the model from the training data based on the current hyperparameter value).

In the case of considering the estimated parameters, the optimized parameters are selected from the verification samples, but the parameters are only estimated from the training data. The idea of verification is the out-of-sample test of the simulation model. Hyperparameter tuning is equivalent to searching for a certain degree of model complexity to produce reliable out-of-sample performance.

The matching of the verification samples is of course not the real out-of-sample, because they are used for tuning, and tuning is in turn an estimated input. Therefore, the third or test sub-sample, which is neither used for estimation nor tuning, is truly out of sample and therefore used to evaluate the predictive performance of the method(Gu et al., 2020).

This study mainly uses a method of random sampling to enlarge the size of negative sample sets into 768 with is the same as the sample size of positive label due to the unbalance class which will introduce in Section 3.3 in detail.

We also test other method which can deal with the unbalance class, and the result will be shown in Section 4. Another important problem for our dataset is the label we use may be not credible, and we also introduce the solution in Section 3.4 and show the results in Section 4. We did not use the validation set because the dataset was too small. Finally, divide the dataset into training set and testing set account for 70% and 30% of the total samples size respectively.

3.2 Model selection

The traditional credit evaluation model is mainly based on the subjective judgment of enterprise credit by professional raters, which is intuitive and easy to operate. However, the model depends on the professional quality of appraisers and has great uncertainty, which affects the effectiveness of the evaluation results (Cai & Yuan, 2005; B. L. Fan &Zhu, 2003) .

Using mathematical statistical model to evaluate enterprise credit is simple and easy to realize, but there are also some obvious deficiencies. Firstly, mathematical statistical models require data to obey empirical assumptions, which is often difficult to meet in reality. Secondly, most mathematical statistical models use financial indicators as the evaluation basis, ignoring the impact of qualitative indicators on credit rating. Finally, the mathematical statistical model cannot dynamically reflect the credit status of enterprises(Chang, 2015).

Machine learning methods have developed rapidly in the field of credit evaluation. The mathematical model is established by analyzing the data, and the data law is found through self-learning, which avoids the limitation of sample data distribution. In addition, machine learning also has a good application in the credit evaluation of small and medium-sized enterprises, because it can deal with noise and incomplete data well(Q. Wang & Yao, 2018).

For the traditional model mainly relying on expert experience, in which all parameters are formulated by experts, it is not used in this study. The mathematical statistical model is represented by Z-score and zeta model. Although it has been widely recognized, it is not suitable for this study, because of necessary financial indicators lacked in the data used in this study and these two models mainly suiting for the credit evaluation of large and medium-sized enterprises and manufacturing enterprises. Therefore, this study focuses on a more flexible and applicable machine learning model.

3.3 Unbalance class

Strictly speaking, any dataset with an uneven class distribution is unbalanced. In general, a dataset is considered unbalanced when there is a significant or, in some cases, extreme mismatch between the number of instances of each category in the dataset.

In other words, a category imbalance occurs when the number of samples representing one class in the dataset is much lower than the number of samples representing other classes. As a result, one or more classes may be underrepresented in the dataset. Since most of the raw data collected in the real world meets this definition, this simple definition has attracted a lot of attention from researchers and practitioners.

Evaluation of loan credit risk of SMEs is a typical situation with unbalance class. There are often much more common enterprises with no default record than the default enterprises. So, we will discuss the problem of unbalanced class and the general methods to deal with it.

3.3.1 Problem ofunbalanced class

Generally, the category with relatively large sample size becomes the majority class, while the category with relatively small sample size becomes the minority class.The direct consequence caused by unbalanced class is that when the sample size gap is large enough, the importance of majority classes in loss function will ‘drown’minority classes, which makes the model more inclined to make the majority classes correctly and ignore the minority classes(Japkowicz & Stephen, 2002).

The number of minority samples in unbalanced datasets is small, so the classifier cannot learn the exact pattern, resulting in low classification accuracy of minority samples. Chawla et al.(2002) points out that when a small number of minority class instances are representative (fixed imbalance ratio), the classification error rate caused by unbalanced class distribution decreases. This is because, despite the uneven class distribution, you can better learn the patterns defined by the positive class examples.

Generally speaking, the goal of supervised learning is to optimize the accuracy of the entire dataset, which may cause the classifier to ignore the performance of each class. In particular, in an unbalanced dataset, if the random classifier predicts all the instances as a majority class, it can achieve very high classification accuracy even if all the few instances are misclassified. Therefore, measures appropriate to the classification of unbalanced datasets should be used.

Data overlap means that instances of two categories overlap to some extent in the feature space, that is, the classification boundary cannot be clearly determined. This makes it more difficult for the classifier to learn the discriminant rules, resulting in the misclassification of minority class samples. The unbalanced data with no overlap in the feature space of the two types of data has little effect on the performance of the traditional classifier.

3.3.2 Deal with unbalanced class

In recent years, the research on the classification of unbalanced datasets has attracted much attention, which has attracted the exploration and research of experts and scholars at home and abroad. For the classification of unbalanced datasets, the discussion and research can be carried out from the aspects of resampling of data, proposal and improvement of classification algorithm, design of evaluation indicators, etc.

Generally, unbalanced data classification can be solved by adjusting the data distribution which includes over-sampling, under-sampling, data generation, etc.This section will introduce the domestic and foreign research status of unbalanced dataset classification task from two aspects: Sample level and Algorithm level.

3.3.2.1 Samplelevel

As for the related algorithms of unbalanced datasets, at the data level, the balanced category distribution is realized mainly by adjusting the category proportion of input datasets, which is called resampling algorithm. These algorithms can be further divided into undersampling, oversampling and mixed sampling algorithms.

The basic undersampling algorithm is Random Under-sampling (RUS), which randomly eliminates a certain number of diverse samples to balance the category proportion of the dataset. The disadvantage of this algorithm is that it is easy to lose important potential information in various samples, which may lead to underfitting problem.

Hart (1968)proposed a Condensed Nearest Neighbor Rule (CNN). If the category of a sample in the dataset is different from the majority of its 3 Nearest Neighbor samples (i.e., at least 2), the sample can be eliminated from the dataset.

Laurikkala(2001)proposed an improved CNN algorithm, Neighborhood Cleaning Rule (NCL), and proposed rules for sample elimination: For each sample in the training set, find 3 samples of its nearest neighbors. If the sample is of most classes and 2 or more of the 3 nearest neighbor samples are different from their classes, we should delete it; if the sample belongs to a minority class and two or more of the three nearest neighbor samples are different from their categories, we should delete the most class of the three nearest neighbor samples.

The performance of NCL algorithm is poor when dealing with data sets that can be overlapped. Based on Consensus Clustering (CC), Onan et al. (2016) runs the selected clustering method repeatedly on the data subset, so as to provide indicators on clustering stability and parameter selection, so as to build a well-balanced data set.

There are three kinds of oversampling algorithms to deal with class imbalance:

1) Random oversampling, focused oversampling and synthetic oversampling. Random over-sampling algorithm (ROS) copies the number of samples in minority classes until the number of samples in minority classes is consistent with the number of samples in a majority class. However, the random over-sampling algorithm exacerbates the overfitting problem of the classifier on the training data set, and if the initial data set has the characteristics of high dimension and multi-noise, these duplicated few class samples will increase the training time of the classifier.

2) Focused over-sampling (FOS) also copies a few class samples, but only those at the boundary between the majority class and the minority class.

The classic representative of Synthetic Oversampling algorithm is the Synthetic Minority Oversampling Technique (SMOTE), which generates Synthetic samples to solve the problem of class balance. For example, SMOTE method proposed by Chawla (2002) combines samples of minority category to generate new samples. It particularly searches similar minority samples by K neighbor and combines in a linear form with random coefficients.

3) Adaptive Synthetic Sampling (ADASYN) based on SMOTE algorithm adjusts minority class samples generated by SMOTE algorithm through the probability distribution of minority classes, so that the generated data set has a better balance. Han et al. proposed a boundary SMOTE algorithm (Borderline SMOTE), which generates synthetic samples on the boundary of the data set but exacerbates the problem of category mixing (Shao et al., 2014).

Naseriparsa et al. (2020)proposed a region-based SMOTE algorithm (RSMOTE), which divided the sample domain into four categories based on the density of a few types of samples: general, semi-general, semi-important and important. The general domain is the region with the highest density, while the important domain is the region with the lowest density. Samples from different regions generate composite samples in different proportions.

For mixed sampling, Songwattanasiri et al. proposed a synthetic Minority over-sampling and under-sampling technique) to solve the class imbalance problem. This algorithm combines SMOTE's over-sampling technique and Reduction Around Ceremoids (RAC)'s under-sampling technique (Songwattanasiri & Sinapiromsaran, 2010). Hussein et al. (2019) proposed an adaptive over-sampling algorithm based on SMOTE, which represented the parameter selection problem in SMOTE algorithm as a multi-objective optimization problem.

In summary, the resampling algorithm is used to reconstruct the unbalanced data set into the balanced data set at the sample level. The traditional classifier can be directly applied to the balanced data set for classifier training, and a better classification effect can be obtained. Next, we will introduce three types of SMOTE oversampling method in detail.

3.3.2.1.1 SMOTE

If there is a significant imbalance in the data, the results of the prediction are often biased, that is, the classification results will be biased towards the more observed categories.How do you deal with this kind of problem? The simplest and most crude way is to construct 1:1 data. Either cut off a part of the larger class (i.e., undersampling), or Bootstrap the smaller class (i.e., oversampling).

But there are problems with this approach. In the first method, the chopping out of the data leads to the loss of some implicit information. In the second method, there is a simple copy of the sample that is put back, which makes the model over-fit again.In order to solve the non-equilibrium problem of data, Chawla proposed SMOTE algorithm in 2002, that is, the synthetic minority oversampling technology, which is an improved scheme based on the random oversampling algorithm.

This technique is a commonly used method to deal with unbalanced data at present, and has been unanimously recognized by the academia and industry. The following is a brief description of the theoretical idea of this algorithm.

The basic idea of SMOTE algorithm is to analyze and simulate a small number of category samples, and add new artificially simulated samples to the dataset, so that the category in the original data will no longer be seriously unbalanced. The simulation process of this algorithm adopts KNN technology, and the steps of simulation generation of new samples are as follows:

1) The nearest neighbor algorithm of sampling is used to calculate the k nearest neighbors of each minority sample.
2) N samples were randomly selected from k nearest neighbors for random linear interpolation.
3) Construct a new minority sample;

4) The new sample is combined with the original data to produce a new training set.

The main problems of the algorithm are as follows:

1) There is blindness in the selection of k nearest neighbor, and the selection of k value is subjective, and its lower limit is limited by the sampling rate *N*. The sampling rate *N*can be obtained by the number of positive and negative samples in the experiment. However, the upper limit of k value is really not conditional, and can only be tried through experiments.
2) SMOTE does not change the data distribution of the unbalanced dataset, and the randomness of the synthesized region is likely to lead to the marginalization of the distribution. The distribution of samples of minority classes determines the selection of its neighborhood. If the sampling point is at the edge of the majority class and the minority class, then the sample points synthesized under the rules will become more and more marginalized. The blurring of classification boundary will increase the difficulty of classification.

3.3.2.1.2 Borderline SMOTE and SVM SMOTE

Borderline SMOTE algorithm is an adaptive sampling algorithm of the basic SMOTE algorithm. The traditional SMOTE algorithm adopts linear interpolation and lacks consideration of the distribution characteristics of sample sets. In order to solve the limitation of the marginalization of synthetic samples in the above problem, the distribution of samples was considered when selecting the set of seed sampling points.

Most classification algorithms tend to have a well-defined decision boundary, meaning that the classification model will distinguish each category boundary point as accurately as possible. Based on the above analysis, points far away from the classification boundary have relatively low impact on the classifier, so Borderline SMOTE algorithm is proposed. The algorithm steps are as follows:

1) For each sample point $x_i$in the sample set P of the minority class, find its nearest M sample points in the whole dataset to obtain the number of sample points m_i of the minority class.
2) If m_i=0, that is, most class sample points surround the point x_i, then this point is determined to be a noise point, and no operation is carried out. If $\boldsymbol{m_i > \frac{m}{2}}$, i.e., $\boldsymbol{x_i}$ is surrounded by more than half of the sample points of minority classes, so it is not a boundary point and does not operate. If $\boldsymbol{0 < m_i < \frac{m}{2}}$, i.e., most of the sample points of the majority class are around x_i, so we can temporarily identify the point where

x_i is at the boundary position, and put this point into a set Danger.

3) For each sample point in Danger, SMOTE algorithm is used to generate new sample points.

Compared with Borderline SMOTE, SVM SMOTE just replace the method which determines classification boundary with SVM model. More detail in Mathew et al. (2014).

3.3.2.1.3 SMOTE tomeklinks

Tomek Links was proposed by Van Gitomek in 1976. It is an effective data cleaning technology. With the wide use of sampling technology in unbalanced classification problems, Tomek links technology is often used to solve the repeated item problems in the process of sampling.Its simple algorithm idea is as follows:

(1) Suppose there is a sample pair $\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right)$, denote the Euclidean distance between the sample$\boldsymbol{x_i}$and $\boldsymbol{x_j}$by $\boldsymbol{d}\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right)$.

(2) If there is no sample x_k which make such that $\boldsymbol{d}(\boldsymbol{x_i}, \boldsymbol{x_k}) < \boldsymbol{d}\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right)$or $\boldsymbol{d}\left(\boldsymbol{x_j}, \boldsymbol{x_k}\right) < \boldsymbol{d}\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right)$,then it is called Tomek links.

After the use of composite sampling, some duplicate samples appear at the boundaries of different categories, then the ‘noise points’ on these adjacent boundaries can be removed using the Tomek Links method until both samples in the nearest neighbor sample pair are from the same category.

The traditional oversampling is based on random sampling with put back to achieve the purpose of simple copying and adding a few samples, but this method causes data redundancy, which is easy to cause the phenomenon of overfitting. Therefore, SMOTE algorithm to oversample can avoid the overfitting phenomenon caused by a large number of repeated samples in the random sampling back to a certain extent.

Although SMOTE method generates a small number of samples through linear interpolation between two samples, the sample space of a small number of samples expands to the sample space of other categories, which may cause the space originally belonging to the majority of samples to be ‘invaded’ by a small number of samples, and also lead to overfitting of the model. By looking for the Tomek Links pair, you can find the noise points or boundary points. Removing the Tomek Links pair is a good way to solve the ‘intrusion’ problem.

So SMOTE Tomek Links method is just an additional filtering step which is Tomek Links method after oversampling .

3.3.2.2 Algorithm level

Generally speaking, cost sensitive algorithm and ensemble learning algorithm are mainly used to solve the classification problem of unbalanced datasets at the algorithmic level. Cost-sensitive algorithms usually do not change the distribution of the dataset. Ensemble learning algorithm can enhance the classification effect by combining different classifiers.

Traditional classification algorithms perform classification tasks based on balanced datasets. Support vector machines implement classification by learning hyperplanes between categories.

However, the Radial Basis Function (RBF) of SVM is usually selected by experience. And the classification effect will be significantly reduced when the SVM algorithm deals with unbalanced datasets. C. Zhang et al. (2019) proposed a border-resampling Feature Elimination (SVM-BRFE) algorithm based on SVM, which iteratively selects features based on boundary Resampling technology. C. Wang et al. (2020) characterized heterogeneous datasets as a Tensor form, which were classified by a Support Tensor Machine (STM).

Traditional classification algorithms assume that the degree of misclassification of a dataset sample can be formulated as a ‘cost’. Lu et.al(2019)proposed a Cost Matrix (CM) algorithm. For each category, the average cost of its misclassification constituted a matrix. By constructing the cost matrix, these algorithms assign greater weight to the misclassification cases of a minority of class samples (relative to the misclassification weight of the majority of class samples).

For the classification task of different real datasets, the cost of misclassification may vary greatly with different datasets. Shao et al. (2014)proposed an improved weighted Lagrange dual support vector machine algorithm, but this algorithm is sensitive to parameter adjustment and cannot be directly and universally applied to unbalanced datasets. J. Lee and Yoon (2017) proposed a weighting regulator to be applied to SVM, this algorithm can be used as a weak learner of ensemble learning to solve the classification problem of unbalanced datasets.

By combining different types of weak classifiers with specific rules, hybrid algorithms can take advantage of the advantages of each weak classifier to break through the limitations of traditional classifier algorithms. This kind of algorithms can also be called ensemble learning algorithms.

Alam et al. used segmentation techniques to generate balanced data, and then constructed an ensemble classifier to solve the classification and regression tasks (Alam et al., 2018). B. Sun et al. proposed an integrated classification model (Handling Imbalanced Data with

Concept Drift (HIDC)), which can effectively deal with the Concept Drift of Datasets (B. Sun et al., 2018). Y. Liu et al. proposed a set classification framework that combines evolutionary undersampling with feature selection(Y. Liu et al., 2019).

The evolutionary undersampling technique in this framework measures the distribution of unbalanced dataset firstly, and uses evolutionary algorithm to optimize its data distribution to achieve the balance of dataset. Y. Sun et al. (2018)proposed a multi-classifier system based on Bagging technique, which generated multiple balanced datasets through the sample balancing method, and constructed the multi-classifier system from these datasets .

The sample balancing method in the system uses the clustering algorithm to divide the majority of classes into multiple class clusters and combine the majority of class clusters with the minority of classes respectively to generate multiple balanced datasets.

Chawla et al. integrated SMOTE algorithm based on ADABOOST technology. In the iteration of ADABOOST, SMOTE algorithm was used to generate a balanced dataset to train the classifier (Chawla et al., 2002). D. Li et al.(2016) based on particle swarm optimization algorithm, adopted ADABOOST algorithm to solve the problem of multiple classification of unbalanced datasets.

Xionget al.(2016)proposed a KAcBagundersampling method based on the sample weight.The algorithm adopts the integrated study of Bagging thought which calledAdaCost weight updating method.The main idea is: firstly, use K-means algorithm to multiple clustering of datasets and update sample weight (pay attention to the variety of clustering center).Then undersample majority class through the sample weight and train the model with AdaCost method, finally, the final classifier is generated by the weighted voting of the weak classifier.

The application of ensemble learning improves the adaptability and robustness of the original classifier, and has a better classification effect for the classification problem of unbalanced datasets.

However, large datasets are often high-dimensional and noisy. The running of ensemble learning algorithm often takes a long time. Therefore, the ensemble learning algorithm needs to be further studied and optimized when dealing with high-dimensional and multi-noise datasets.

3.4 Label credibility

The application of machine learning methods to the credit evaluation scenarios of small and micro-sized enterprises has an important issue that has not been discussed in detail in the literature, that is, the credibility of labels.

In general, the loan time of the company may be longer, so the trust-breaking company can be labeled according to the repayment situation of the loan due, but for the sample whose payment has not yet been due, the label is not accurate. The data sample interval used in this study is from 2011 to 2016, but the loan term ranges from 1 to 6 years, about 2 years on average.

Therefore, there are a large number of samples of companies marked credible, which actually have quite high default risk. In this case, the features of these ‘false’ common companies may be similar to the feature distribution of faithless companies and are very different from the feature distribution of ‘true’ common companies, ,which will cause great interference to the discriminant and greatly reduce the discriminant efficiency.

Generally, weakly supervised learning methods are usually adopted for this situation. Next, we will introduce the traditional weakly supervised learning methods and then introduce the method we used in this study.

3.4.1 Introduction of weakly supervised learning

Traditional Supervised Classification is one of the most widely studied learning frameworks in machine learning (Blum & Mitchell, 1998). For each object in the real world, the learning system uses a sample (usually a feature vector) to describe the properties of the object, and a category label to describe the semantic information of the object. The learning system predicts the markers of unlabeled samples through learning modeling of known labeled training samples.

Traditional supervised classification methods are based on strong supervised assumptions :(1) Assumption that the number of labeled samples is sufficient; (2) Assume that the labeling information of the labeled samples is correct; (3) It is assumed that the mark of the sample is single, that is, each sample has only one true mark. For the classification problems satisfying the above strong supervision hypothesis, the traditional supervised classification framework has achieved great success .

However, the above strong supervision hypothesis actually simplifies the problem in the real world, and it does not hold true in many real world application scenarios. What we get in real applications is usually weak monitoring information.

Evaluation of Loan Credit Risk of SMEs often meets the incorrected label which is contrary to the second assumption. In recent years, in order to solve the classification problem under weak supervision, researchers have proposed a variety of algorithm frameworks. In particular, in order to solve the problem of inadequate supervised learning, semi-supervised learning and active learning have drawn extensive attention from researchers.

For semi-supervised learning, researchers note that although unlabeled samples have no labeled information for the model to learn, the presence of a large number of unlabeled samples can help the model learn the potential distribution information of the data.

Therefore, semi-supervised learning aims to train the classification model by comprehensively utilizing a small number of labeled samples and a large number of unlabeled samples, so as to obtain a classifier with better performance than the classifier trained by only a small number of labeled samples and make up for the deficiency of labeled samples. Unlike semi-supervised learning, which has no human intervention, active learning assumes an 'oracle', such as a human expert.

The learning system selects the most valuable unlabeled samples (or sample labeled pairs) for expert labeling by designing effective sample (or sample labeled pair) selection strategies, and queries the truth markers of unlabeled samples(Settles, 2012).Therefore, the focus of active learning is to design an effective sample selection strategy, aiming to obtain a better performance classifier by using fewer but high-value labeled samples.

In other words, under the same cost (such as the same number of samples), active learning can obtain higher classification accuracy by selecting more valuable samples than randomly selected samples.

In order to solve the problem of inaccurate supervised learning, researchers proposed a biased label learning framework (Cour et al., 2011). In the biased tag learning framework, each object is given one or more tags, but only one of the tags is true, and the rest are noise tags. Partial label learning aims to learn under the weak supervised information containing noise labels and obtain a classifier that is robust to noise labels.

In order to solve the problem of polysemous supervision, that is, the problem of excessive output space, multi-label learning has been extensively studied by researchers(Bhatia et al., 2015). Most multi-label learning algorithms assist the training of classifiers by making full use of the correlation between markers.

In addition, in order to solve the problem of insufficient supervised learning in more complex polysemous supervised scenarios and inaccurate supervised learning in polysemous supervised scenarios, researchers proposed multi-label active learning framework and partial multi-label active learning framework respectively. Among them, multi-label active learning designs effective sample (or sample label pair) selection strategies by combining multi-label learning and active learning (Qing&Xin, 2015).

Different from the hypothesis of partial marker learning, partial multiple marker learning assumes that there can be multiple true markers among multiple candidate markers in each sample.

We mainly focus on the biased label learning because this type of method is suitable for our situation. Traditional biased label learning framework is an important weakly supervised learning framework. In this framework, each sample is associated with a set of candidate markers in the tag space, but only one of them is the true marker of the sample.

In recent years, researchers have proposed many biased label learning algorithms. Among them, some methods assume that each marker is equally important, and then train the classification model by averaging the models of all candidate markers with labeled samples to predict the markers of unlabeled samples.

In particular, BeringerandHüllermeier(2006) used the candidate markers of the neighborhood samples of the unlabeled sample to vote, and the marker with the most votes was the predictive marker of the unlabeled sample. Cour et al. (2011) trained the classification model by averaging the model output of all candidate markers with labeled samples to estimate the experience loss of the classification model.

In addition, some methods set up a parameterized classification model for each marker, and regard the true real index of the labeled sample as a latent variable to be optimized. The potential true markers and model parameters are optimized by iteratively optimizing the objective function based on maximum likelihood estimation or the objective function based on maximum interval (Nguyen & Caruana, 2008) through a process of Expect-Maximization (EM).

In order to further improve the effect of Partial Label learning, M.L.Zhang & Yu (2015) proposed a two-stage Partial Label Learning method (IPAL). The method first constructs a k-nearest neighbor graph based on all labeled samples, and then performs label propagation on the constructed graph to identify the true markers of labeled samples.

Then, based on the identified true markers, the K-nearest neighbor voting method is used to predict the markers of unlabeled samples. In addition, unlike IPAL, which explicitly identifies real markers in labeled samples in the first phase, M. L. Zhang et al. (2016) proposed a new two-stage Partial Label Learning via feature-aware disambiguation (PL-leaf) based on the local topology of the Feature space of the sample.

In the first stage, the method estimates the label confidence of each candidate marker with a labeled sample by forcing the manifold structure of the sample in the label space to be consistent with that of the sample in the feature space. In the second stage, the method trains a regularized multi-output regression model to predict the markers of unlabeled samples based on the labeling confidence generated in the first stage.

The above methods are all biased label learning algorithms based on disambiguation, and the index with the largest output value of the model is denoted as the true index of the sample. By extending the Error-Correcting Output Codes (ECOC) approach, X. L. Zhang (2014) proposed a novel single-tag, Error-Correcting Output Codes learning algorithm based on non-disambiguation strategies. In the coding stage, the method first generates a random binary (-1, +1) coding matrix, and then divides the tag space into two disjoint subsets based on each column of the coding matrix to construct a binary classifier.

In the decoding stage, the codeword of each sample is obtained based on the output of the sample in each binary classifier, and the distance between the codeword and the codeword of each marking category in the encoding stage is calculated. The category mark closest to the codeword is used as the output mark of the test sample.

All the above partial label learning methods restrict each sample to have at most one true label. In the multi-tag learning scenario, each sample can have multiple tags at the same time. It is obvious that biased multiple marker learning is more challenging than traditional biased marker learning, which is a special case of biased multiple marker learning. In addition, due to the rapid expansion of the marker space in the multi-marker scenario, it is necessary to consider the marker relevance in the multi-marker learning framework to enrich the supervisory information.

In order to solve the problem of Partial label Learning in multi-label Learning scenarios, some research proposed a Partial Multi-label Learning (PML) framework. The PML framework consists of two algorithms: PML-LC (Partial Multi-Label Learning with Label Correlations) and PML-FP (Partial Multi-Label Learning with Feature Prototypes). PML assumes that each candidate marker has a confidence value, indicating the reliability of the candidate marker as a true marker, and then optimizes the marker confidence and training classification model by minimizing the ranking loss of the confidence-weighted markers.

In particular, PML-LC and PML-FP introduce an additional constraint on the basis of PML based on marker association and feature prototype respectively to jointly optimize the marker confidence and prevent model overfitting.

However, the framework not only needs to train an independent parameter matrix for each marker to predict sample markers, but also needs to optimize the ordering loss of a large number of paired markers. Therefore, with the rapid expansion of the tag space, the computational efficiency of this algorithm framework will decrease significantly.

3.4.2 Tri-Training method

We will introduce a Tri-Training method proposed by Zhou & Li (2005)which will be used in the next section. This method is a typical type of weakly supervised learning method which is extended from Self-Training method and Co-Training method.

Self-training (also known as self-teaching)(Raina et al. 2007) is the simplest and most commonly used semi-supervised learning algorithm.

The general steps of self-training are as follows: in the first step, the labeled sample set is used to train to get the initial classifier. The second step uses the acquired initial classifier to mark the unlabeled samples. The third step is to add some labeled samples with high confidence in the results to the labeled sample set. The fourth step is to re-train with the updated training sample set to obtain a new classifier.

The whole process is iterated until the algorithm reaches a certain convergence or stop condition. Self-training algorithm has been widely applied to face recognition, handwritten word recognition, text classification, image classification, intrusion detection and many other fields. The basic assumption of self-training is that when the classifier pre-estimates the samples, the samples with high confidence are more likely to be accurately classified.

For example, when SVM classifies samples, those samples far away from the classification interface are generally considered to be correctly classified samples. Based on this assumption, self-training is very simple to handle. Suppose there are two datasets A and B, where A is labeled data. And B is unlabeled data. The methods of self-training are as follows:

- Train A classification model M from labeled dataset A.
- Use the model to predict dataset B.
- Add K samples with high confidence in the prediction results, together with their labels, into training dataset A, and delete them from dataset B.
- Go back to the first step until some convergence or stop condition is reached.

A big disadvantage of self-training is that if a sample is misclassified, it will be added to the original Training set. In the subsequent Training process, the misclassified sample will only make more and more mistakes and may lead to mistakes of other samples. As a result, the classification performance of the classifier is reduced.

Co-training(Hady & Schwenker, 2008) is also known as collaborative Training. The original collaborative Training algorithm was proposed by Blum and Mitchell (1998) . In literature, they assume that the dataset has two fully redundant views.

The first view is that each attribute set is sufficient to describe the problem, that is, in the case of a large number of training samples, there is enough to learn a strong learner on each attribute set. Second view: Each attribute set is conditionally independent of the other, given the tag. Given the existence of these two fully redundant views, they concluded that the requirement of fully redundant views could be met for many tasks.

Zhou and Li (2005) proposed an improved algorithm of cooperative Training algorithm, Tri-Training , which changed the two classifiers adopted in the co-training algorithm to adopt three classifiers for Training to improve learning performance. This would no longer follow the two fully redundant views assumed by Blum andMitchell(1998).

In the process of constructing the model, Tri-Training algorithm adopts two of the classifiers to predict and label the unlabeled samples, and adopts the consensus method to vote to determine the predicted classification results. If the voting results of the two classifiers are consistent, and the confidence of the marker is higher than the set threshold, then the unlabeled sample together with the labeled category will be added to the training set of the third classifier for training. If the results of the two classifier votes do not agree, then the training is repeated until no updates are made or the iteration is completed.

Compared with the co-training algorithm, the Tri-Training algorithm has the following advantages. First, the sample space does not need to meet the assumption of two fully redundant views. Second, there are no constraints on the supervised learning algorithm. Therefore, Tri-training algorithm is superior to co-training algorithm, and its application is more extensive than that of co-training.

Later, they further extended on the basis of the three classifiers and proposed a co-forest algorithm that could better play the role of ensemble learning. Co-forest algorithm uses a large number of classifiers, and adopts ensemble learning algorithm to combine multiple classifiers, so that its performance is better(M. Li & Zhou, 2007). The core algorithm mainly used in this study is the Tri-Training algorithm, so we will introduce the Tri-Training algorithm in detail below.

a) Choose 3 learners named a,b,c. The training samples were labeled sample L and unlabeled or maybe incorrect sample U.

b) Based on L, the three learners were trained for the first time. Then three learners are used to predict U, and samples with the same prediction by B and C learners are defined as $\boldsymbol{L_a^1, L_b^1, L_c^1}$

c) For learner a, if constraint is satisfied, re-training based on sample set $\boldsymbol{L_1} = \{\boldsymbol{L_a^1}, \ \boldsymbol{L}\}$.Then the same for learner b and c.

d) Repeat above step untildoes not satisfy the constraint equation :

$$\mathbf{0} < \frac{\hat{e}_i^t}{\hat{e}_i^{t-1}} < \frac{|L_{t-1}|}{|L_t|} < \mathbf{1} \tag{3.1}$$

Where $|L_t|$represent the sample size for t-1'th iteration of training and $\hat{e}_i^t$is the prediction error rate of learner. The constraint is obtained by following method simply.

From the PAC theorem, we have equation :

$$\mathbf{P}\left(\mathbf{d}\left(H_i, H^*\right) \geq \varepsilon\right) \leq \delta \tag{3.2}$$

d(.) represents the difference between the two hypotheses,$\boldsymbol{\delta}$ denotes$\boldsymbol{H_i}$the confidence of the hypothesis, H^* represents the true hypothesis, which is equation :

$$m \geq \frac{2}{\varepsilon^2 (1-2\mu)^2} \ln\left(\frac{2N}{\delta}\right) \tag{3.3}$$

Then, we have equation :

$$m = \frac{c}{\varepsilon^2 (1-2\mu)^2} \tag{3.4}$$

Where μ represents the upper limit of noise rate of the classifier, N represents the number of hypothesis spaces, and M represents the sample size. Therefore, for an iteration of the model, it is necessary to satisfy the requirement that the ε of the trainer becomes smaller under the condition of constant δ , means that $\varepsilon_t < \varepsilon_{t+1}$ .Then, we have equation :

$$m^t \left(1-2\mu^t\right)^2 > m^{t-1} \left(1-2\mu^{t-1}\right)^2 \tag{3.5}$$

Let $\boldsymbol{\mu^t}$ represents the noise rate of classification, $\hat{e}_i^t$ is the prediction error rate of learner, $|L_t|$ represent the sample size for t-1'th iteration of training, then, it can be written as equation:

$$\mu^t = \frac{\mu_L |L| + \hat{e}_i^t |L_t|}{|L_t \mathrm{U} L|} \tag{3.6}$$

Substituting μ^t into the above equation, we can get the constraint.

3.5 Evaluationmethod

3.5.1 Cross validation

In order to measure the variance of the evaluation index, it is necessary to conduct several experiments. In order to make the experiment more statistically significant, the experiment design of this subject is carried out by means of cross-validation.

Cross-validation is one of the commonly used precision measurement methods(Bengio & Grandvalet, 2004). In this study, K-fold Cross-validation Method, which is the most common one, is mainly used. The process is expressed as follows. Firstly, the training set is divided into K equal parts. Secondly, K-1 sets are trained to obtain a training model, and the remaining set is used to test the training model.

Then, the process is repeated K times, and the average value of test errors is treated as the final error. Especially, if K=N, the K-fold cross-validation method can also be called Left One Method.

The process is to first take out the *i*th sample as test sample, train the remaining n-1 training sample to get the training model and test the *i*th sample, then put it back. Each time take different test sample until each sample has been taken out.

Considering the efficiency of model fitting, this study uses 10-fold cross validation to validate the dataset as Figure 3.1 Panel (A). Figure 3.2 shows the confusion matrix, in which the predicted value and actual value are both 1 to obtain the corresponding TP value. When both are taken as 0, the TN value can be obtained. Generally speaking, the higher the TP value and TN value, the higher the accuracy of classification.

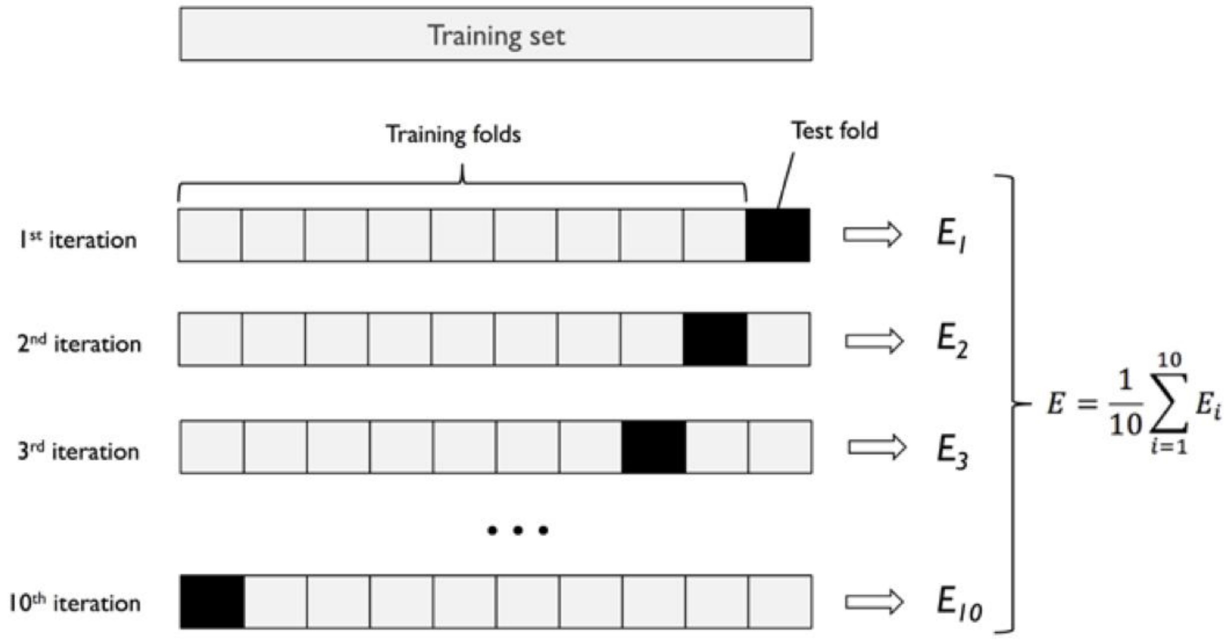
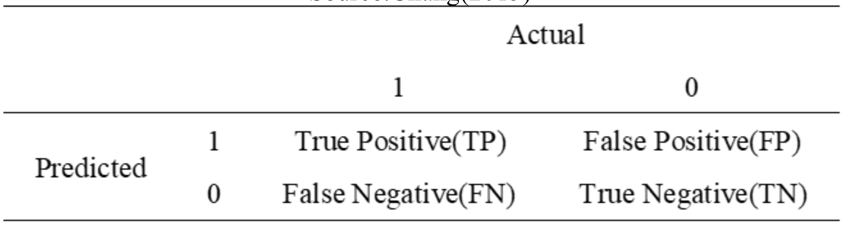


Figure 3.1 10-fold cross validation

Source:Chang(2015)



		Actual	
		1	0
Predicted	1	True Positive(TP)	False Positive(FP)
	0	False Negative(FN)	True Negative(TN)

Figure 3.2 Confusion matrix

Source: Chang(2015)

3.5.2 Evaluation indicator

In order to introduce the use of evaluation indicators in this study, the confusion matrix must be introduced (Bradley, 1997). Consider the situation that the default sample is 1 and the normal loan repayment sample is 0. ‘Predicted’ represents the predicted value of the model for the sample category, and ‘Actual’ represents the actual value of the sample. The confusion matrix is shown as Figure 3.1 Panel (B).

Then, the classical evaluation indicator used in Machine learning will be introduced (Bradley 1997). Accuracy is definedas the total true rate of the model prediction regardless of the normal loan repayment sample and default sample. Precision rate pays more attention to the sample that the model predicts as 1 and recall rate pays more attention to the sample that the actual defined as 0. F1 score is harmonic average of precision rate and recall rate. ROC is defined as the relationship between TPR and FPR.

AUC is the area under the ROC curve is between 0 and 1, which can be used as a numerical value to directly evaluate the classifier. If two ROC curves do not intersect, we can determine the curve which is closest to the upper left corner represents the best performance of the learner. However, the situation is more complicated in practice. If two ROC curves cross, it is difficult to generalize who is superior and who is inferior.

In many practical applications, we often want to distinguish the performance of the learners with similar situation. Then, the AUC area is introduced to deal with it.When comparing the learners, if the ROC curve of one learner is completely ‘enveloped’ by the curve of another, it can be asserted that the performance of the latter is better than the former. If the ROC curves of two learners cross, the more reasonable judgment basis is to compare the area under the ROC curve, namely, AUC

PR curve is similar to the ROC curve, just replace FPR with R. If the PR curve A of one of the learners completely covers the PR curve of the other learner B, it can be asserted that the performance of A is better than B.

- Accuracy =(TP+TN)/(FP+TP+FN+TN):
- Precision rate (P, TPR) = TP/(TP+FP).
- Recall rate (R) = TP/(TP+FN).
- False Positive Rate (FPR)= FP/ (FP+TN)
    - F1 score: Harmonic average of precision rate and recall rate， $2\times\frac{\boldsymbol{P}\times\boldsymbol{R}}{\boldsymbol{P}+\boldsymbol{R}}$.
- ROC curve: the abscissa is FPR and the ordinate is TPR, the ROC curve can be understood as the relationship between the correct and false probability in the positive class at different thresholds.
- AUC: the area of the ROC curve (the intuitive meaning of AUC is to take one positive and negative sample arbitrarily, and the probability that the positive sample score is greater than the negative sample).
- PR curve (precision rate - recall rate curve): the abscissa is R and the ordinate is P.

3.6 Hardware and software used

Machine learning requires a good hardware environment. But the configuration requirements required to train the data using machine learning vary with the scenario.

If the training data is small and the algorithm is less complex, then the required hardware base is not very high. If the training data is very large, then a sufficiently large memory environment is needed. If the algorithm is very complex, then a CPU core with sufficient computing power is needed for non-deep learning algorithm. And if it is a deep learning algorithm, a GPU with sufficient computing power is also needed.

The hardware and software we used are shownin Table 3.2:

Table 3.2 The hardwares and softwares

	Type	Num	frequency	memory
CPU	Ryzen 7 3700X	2	3.6-4.4GHz	
GPU	RTX2070 SUPER	1		6G
Memory				32G
Operating System	Windows 10			
Code	Python 3.6			
Library	Sklearn 0.24			

[This page is deliberately left blank.]

Chapter 4: Evaluation of Loan Credit Risk of SMEs

4.1 Results of the models applications

4.1.1 Logistic regression

This study constructs a Logistic regression model based on enterprises' credit data and use it to predict credit default. The results are as follows:

The empirical distribution diagram of the degree of confidence, which is predicted by the Logistic regression model trained by training set, is shown in the top of Figure 4.1.

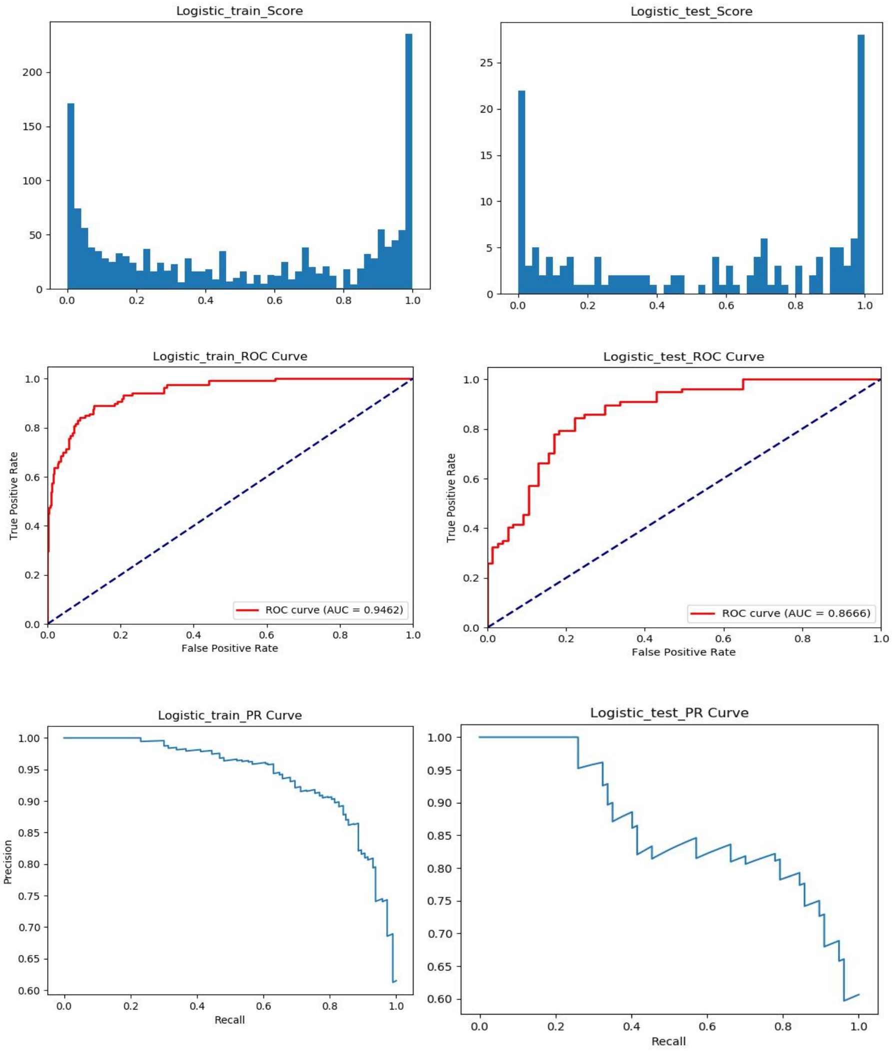


Figure 4.1 Result of 10-fold cross validation (by logistic regression)

The model outputs results in the training set and testing set of ROC (The Corresponding

Receiver Operating Characteristic) and AUC (Area Under Curve) are shown in the middle part of Figure 4.1.

The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.1.

In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.1.

Table 4.1 Result of 10-fold cross validation (bylogisticregression)

	Train					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.869	0.872	0.864	0.868	4.535	0.941
1	0.871	0.881	0.857	0.869	4.454	0.946
2	0.862	0.871	0.849	0.860	4.778	0.940
3	0.862	0.873	0.849	0.860	4.753	0.942
4	0.863	0.873	0.850	0.861	4.728	0.941
5	0.867	0.882	0.849	0.865	4.579	0.946
6	0.865	0.879	0.847	0.863	4.653	0.941
7	0.860	0.875	0.840	0.857	4.828	0.938
8	0.869	0.893	0.839	0.865	4.529	0.944
9	0.869	0.884	0.849	0.866	4.529	0.946
Ave	0.866	0.878	0.849	0.863	4.637	0.943
	Test					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.878	0.968	0.782	0.865	4.207	0.932
1	0.786	0.797	0.766	0.781	7.401	0.888
2	0.883	0.915	0.844	0.878	4.037	0.957
3	0.825	0.805	0.857	0.830	6.056	0.916
4	0.877	0.914	0.831	0.871	4.261	0.944
5	0.786	0.756	0.844	0.798	7.401	0.883
6	0.857	0.848	0.870	0.859	4.934	0.944
7	0.883	0.928	0.831	0.877	4.037	0.963
8	0.838	0.783	0.935	0.852	5.607	0.894
9	0.805	0.783	0.844	0.813	6.728	0.867
Ave	0.842	0.850	0.840	0.842	5.467	0.919

Caption: TP (True Positive) is the number of correct prediction of positive samples. FP (False Positive) is the number of incorrect prediction of positive samples, *i.e.* predicting negative sample as positive one. TN (True Negative) is the number of correct prediction of negative samples. FN (False Negative) is the number of incorrect prediction of positive samples, *i.e.* predicting positive sample as negative one. Precision is the ratio of TP to the sum of TP and FP,*i.e.* the ratio of true positive sample to all positive samples found by the classifier. Recall is the ratio of TP to the sum of TP and FN,*i.e.* the ratio of true positive sample found by the classifier to all positive samples. Accuracy is the ratio of the sum of TP and TN to the sum of TP, FP, TN and FN, *i.e.* the ratio of correct predictions in all predictions. Accuracy can be seen as an overall judgement of the classifier. F1 is the harmonic mean of Accuracy and Recall, which can be seen as a comprehensive indicator of Accuracy and Recall. Loss Value is the final value of the Loss Function, the lower it is, the lower the level of miscalculation value is.

4.1.2 Decision tree

Moreover, this study constructs a Decision Tree model to predict credit default by training it with enterprises' credit data.

First, all training sets are randomly divided into training set and testing set according to the 7:3 ratio. The empirical distribution diagram of the degree of confidence, which is predicted by the Decision Tree model trained by training set, is shown in the middle part of Figure 4.2. The model outputs results in the training set and testing set of ROC (The Receiver Operating Characteristic) and AUC (Area Under Curve) is shown in the top of Figure 4.2. The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.2.

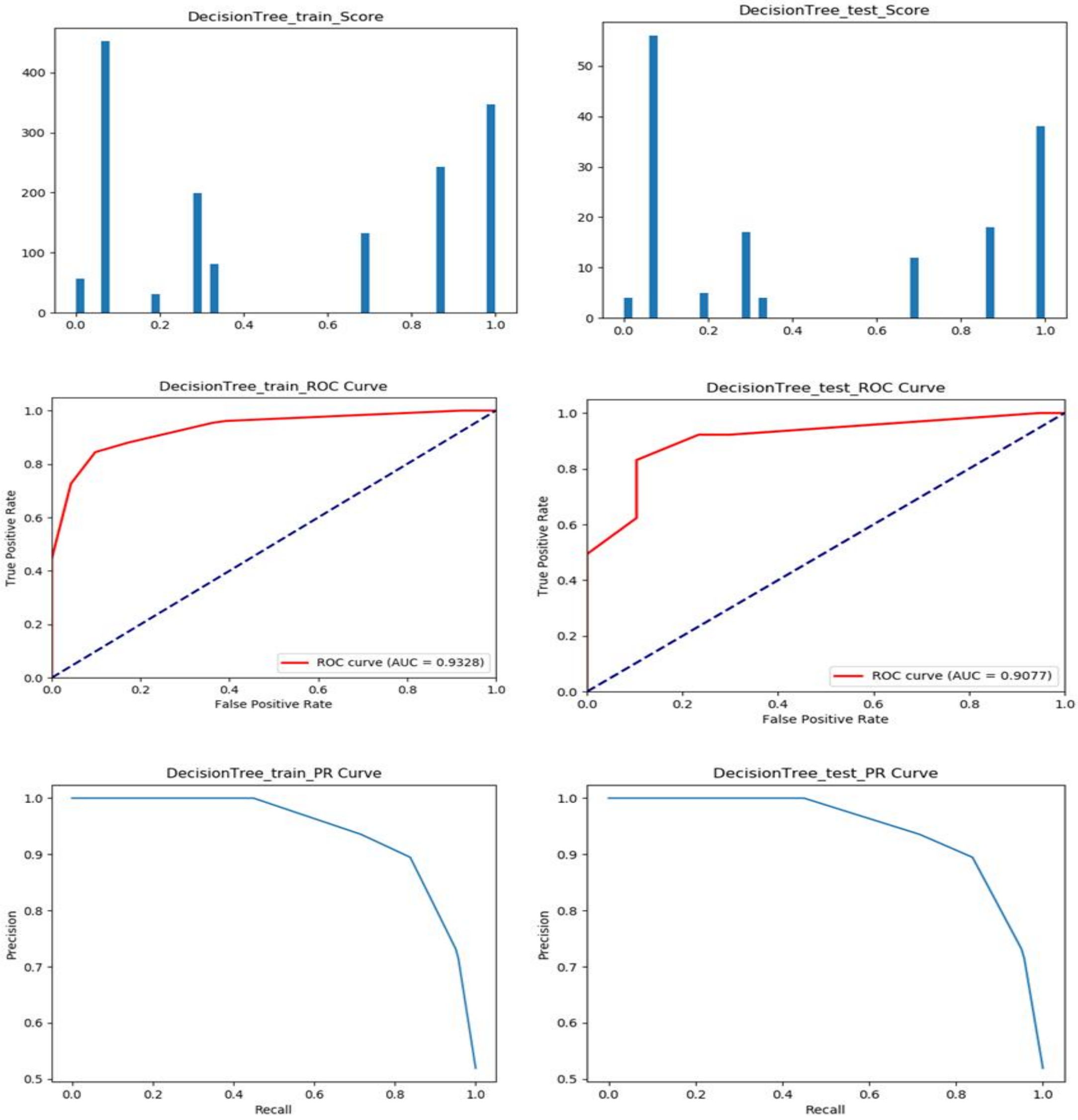


Figure 4.2 Result of 10-fold cross validation (by decision tree)

In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.2.

Table 4.2 Result of 10-fold cross validation (by decision tree)

			Train			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.867	0.853	0.887	0.870	4.585	0.921
1	0.872	0.893	0.846	0.869	4.404	0.937
2	0.869	0.901	0.829	0.863	4.529	0.930
3	0.863	0.895	0.823	0.857	4.728	0.938
4	0.867	0.935	0.790	0.856	4.579	0.927
5	0.889	0.892	0.885	0.889	3.832	0.945
6	0.875	0.898	0.847	0.872	4.305	0.935
7	0.867	0.881	0.850	0.865	4.579	0.918
8	0.870	0.887	0.849	0.867	4.479	0.929
9	0.873	0.896	0.844	0.869	4.380	0.933
Ave	0.871	0.893	0.845	0.868	4.440	0.931
			Test			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.821	0.805	0.846	0.825	6.199	0.891
1	0.857	0.877	0.831	0.853	4.934	0.885
2	0.812	0.808	0.818	0.813	6.504	0.899
3	0.870	0.880	0.857	0.868	4.486	0.934
4	0.857	0.937	0.766	0.843	4.934	0.914
5	0.779	0.742	0.857	0.795	7.626	0.879
6	0.851	0.875	0.818	0.846	5.158	0.926
7	0.877	0.863	0.896	0.879	4.261	0.936
8	0.844	0.827	0.870	0.848	5.383	0.918
9	0.838	0.882	0.779	0.828	5.607	0.908
Ave	0.840	0.849	0.834	0.840	5.509	0.909

According to the results, it can be found that the difference between the test set and the training set is not large, and the PR curve and the distribution of the score and the ROC curve are similar, indicating that there is no significant over-fitting phenomenon. From the results of cross-validation, there is also no significant over-fitting phenomenon, and the effect of the model in the test set is less fluctuating, indicating that the model is more stable in this scenario.

4.1.3 Random forest

This study also constructs a Random Forest model trained by enterprises' credit data, and uses it to predict credit defaults.

The empirical distribution diagram of the degree of confidence, which is predicted by the Random Forest model trained by training set, is shown in the top of Figure 4.3. The model outputs results in the training set and testing set of ROC (The Corresponding Receiver Operating Characteristic) and AUC (Area Under Curve) is shown in the middle part of Figure 4.3.

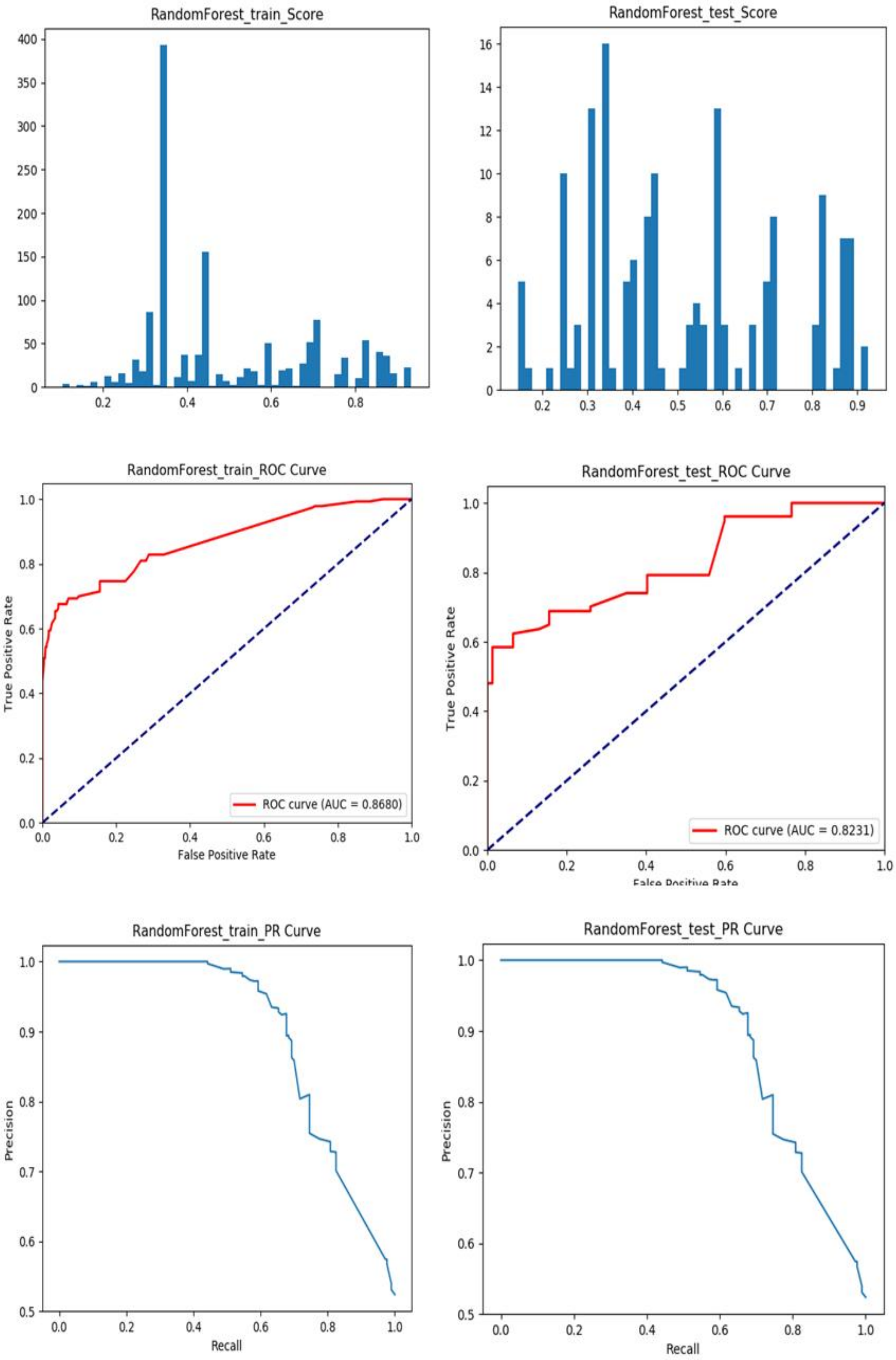


Figure 4.3 Result of 10-fold cross validation (by random forest)

The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.3. In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.3.

Table 4.3 Result of 10-fold cross validation (by random forest)

	Train					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.802	0.910	0.671	0.772	6.828	0.893
1	0.844	0.904	0.769	0.831	5.400	0.919
2	0.816	0.872	0.739	0.800	6.370	0.876
3	0.824	0.891	0.739	0.808	6.072	0.888
4	0.822	0.907	0.718	0.801	6.146	0.883
5	0.814	0.929	0.680	0.785	6.420	0.877
6	0.812	0.912	0.690	0.786	6.495	0.895
7	0.828	0.896	0.742	0.812	5.947	0.900
8	0.832	0.932	0.716	0.810	5.798	0.880
9	0.809	0.902	0.693	0.784	6.594	0.868
Ave	0.820	0.906	0.716	0.799	6.207	0.888
	Test					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.827	0.840	0.808	0.824	5.978	0.890
1	0.812	0.853	0.753	0.800	6.504	0.870
2	0.870	0.938	0.792	0.859	4.486	0.903
3	0.792	0.846	0.714	0.775	7.177	0.878
4	0.844	0.949	0.727	0.824	5.383	0.878
5	0.812	0.914	0.688	0.785	6.504	0.878
6	0.799	0.859	0.714	0.780	6.953	0.881
7	0.844	0.965	0.714	0.821	5.383	0.903
8	0.747	0.757	0.727	0.742	8.747	0.806
9	0.714	0.726	0.688	0.707	9.868	0.823
Ave	0.806	0.865	0.733	0.792	6.698	0.871

According to the results, it can be found that the difference between the testing set and the training set is not large, and the PR curve and the distribution of the score and the ROC curve are similar, indicating that there is no significant over-fitting phenomenon.

From the results of cross-validation, there is also no significant over-fitting phenomenon, and the effect of the model in the testing set is less fluctuating, indicating that the model is more stable in this scenario.

4.1.4ADABOOST

This study also constructs anADABOOST model trained by enterprises' credit data, and uses it to predict credit defaults.

The empirical distribution diagram of the degree of confidence, which is predicted by theADABOOST model trained by training set, is shown in the top of Figure 4.4.

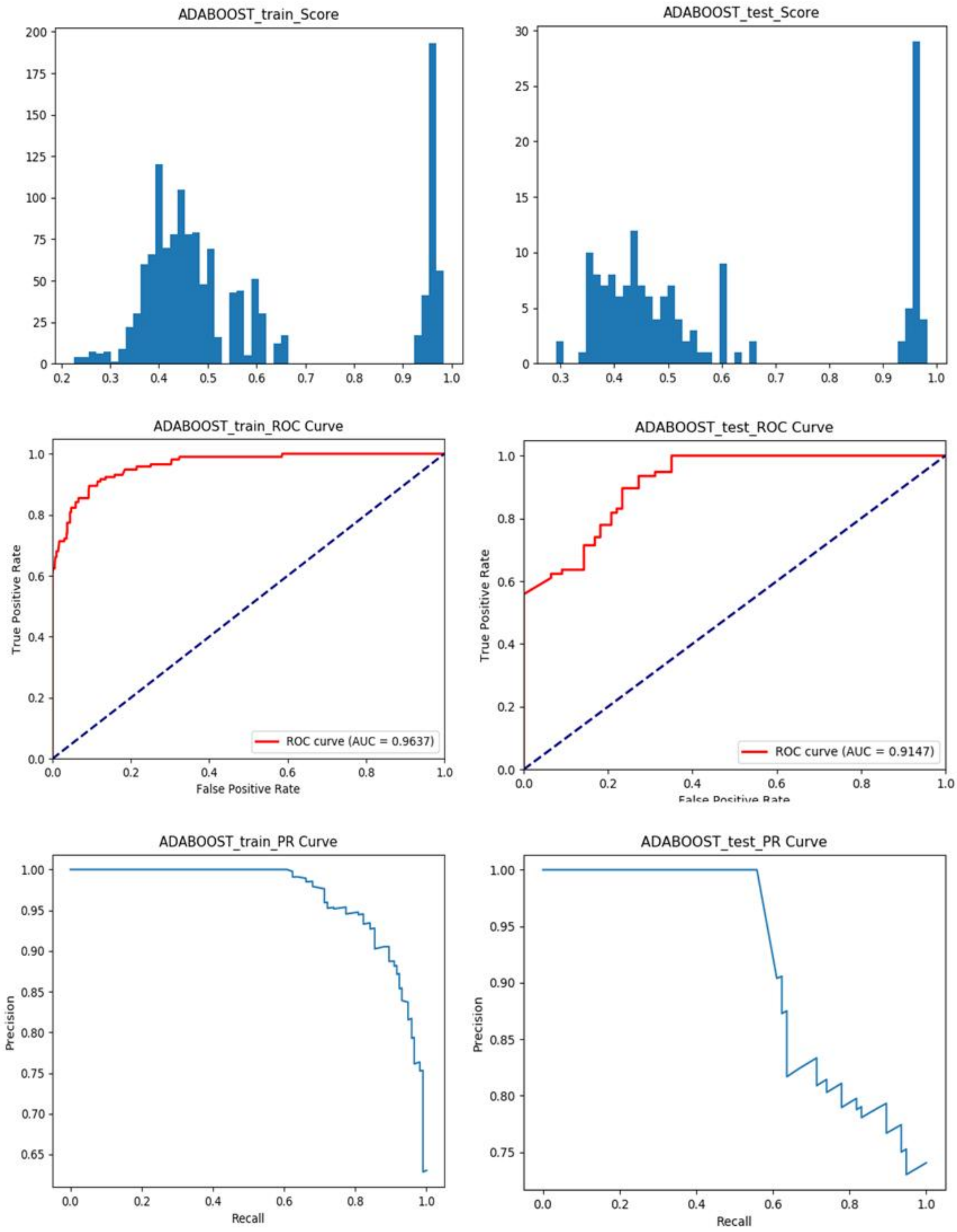


Figure 4.4 Result of 10-fold cross validation (byADABOOST)

The model outputs results in the training set and testing set of ROC (The Corresponding Receiver Operating Characteristic) and AUC (Area Under Curve) is shown in the middle part of Figure 4.4.

The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.4.

In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.4.

Table 4.4 Result of 10-fold cross validation (byADABOOST)

			Train			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.867	0.915	0.808	0.858	4.610	0.955
1	0.834	0.886	0.765	0.821	5.748	0.943
2	0.862	0.924	0.788	0.851	4.778	0.950
3	0.851	0.904	0.785	0.840	5.151	0.946
4	0.840	0.899	0.767	0.827	5.524	0.947
5	0.843	0.903	0.768	0.830	5.425	0.956
6	0.833	0.907	0.742	0.816	5.773	0.947
7	0.854	0.895	0.803	0.847	5.027	0.942
8	0.844	0.912	0.762	0.830	5.375	0.947
9	0.882	0.948	0.808	0.872	4.081	0.964
Ave	0.851	0.909	0.780	0.839	5.149	0.950
			Test			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.872	0.939	0.795	0.861	4.428	0.954
1	0.825	0.847	0.792	0.819	6.056	0.914
2	0.883	0.954	0.805	0.873	4.037	0.946
3	0.812	0.816	0.805	0.810	6.504	0.934
4	0.857	0.923	0.779	0.845	4.934	0.958
5	0.818	0.866	0.753	0.806	6.280	0.928
6	0.786	0.824	0.727	0.772	7.401	0.910
7	0.870	0.938	0.792	0.859	4.486	0.960
8	0.805	0.813	0.792	0.803	6.728	0.937
9	0.786	0.814	0.740	0.776	7.401	0.915
Ave	0.831	0.873	0.778	0.822	5.826	0.935

According to the results, it can be found that the difference between the testing set and the training set is not large, and the PR curve and the distribution of the score and the ROC curve are similar, indicating that there is no significant over-fitting phenomenon.

From the results of cross-validation, there is also no significant over-fitting phenomenon, and the effect of the model in the testing set is less fluctuating, indicating that the model is more stable in this scenario.

4.1.5 XGBOOST

This study also constructs a XGBOOST model trained by enterprises' credit data, and uses it to predict credit defaults.

The empirical distribution diagram of the degree of confidence, which is predicted by the XGBOOST model trained by training set, is shown in the top of Figure 4.5. The model outputs results in the training set and testing set of ROC (The Corresponding Receiver Operating Characteristic) and AUC (Area Under Curve) is shown in the middle part of Figure

4.5.

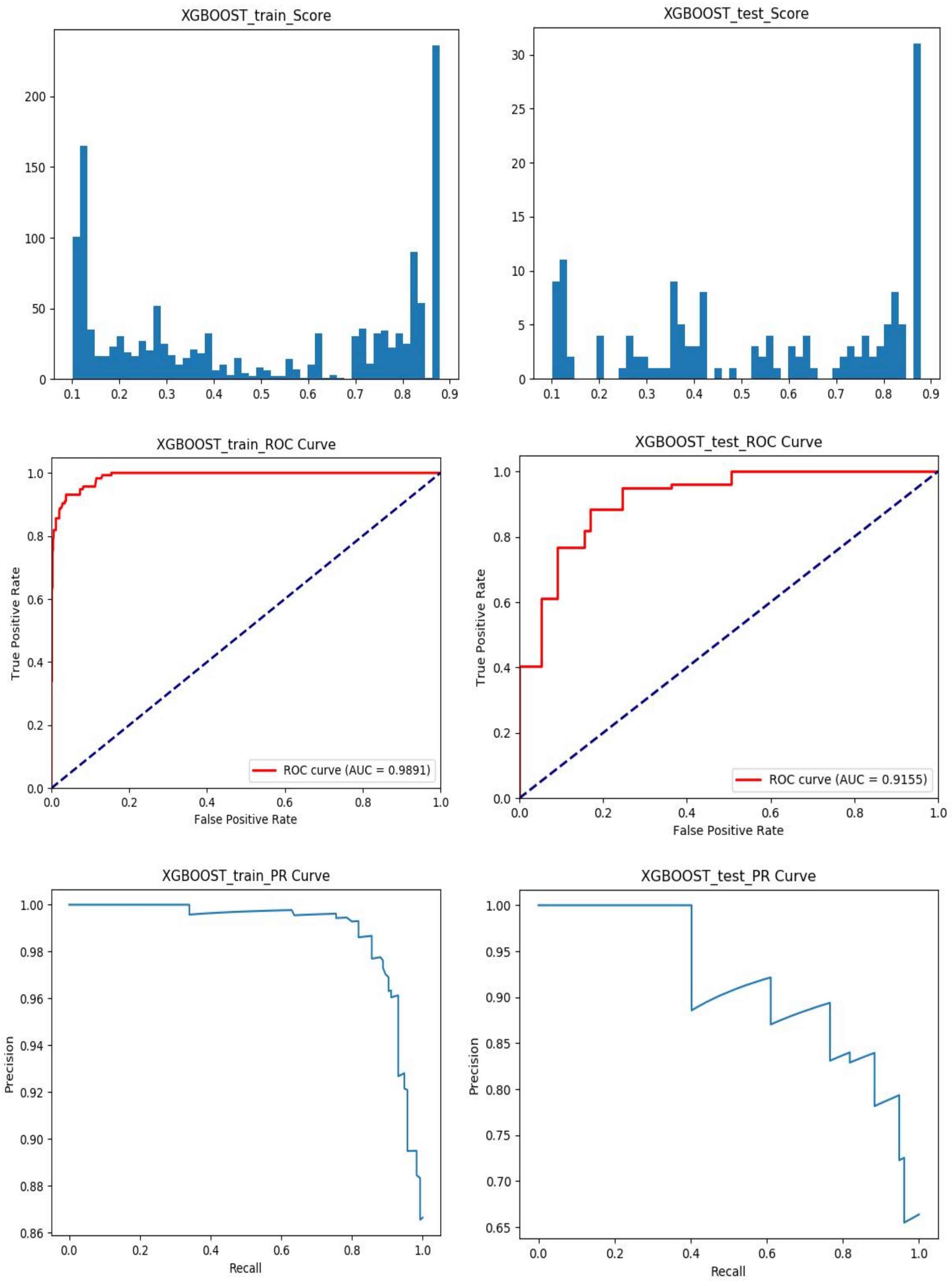


Figure 4.5 Result of 10-fold cross validation (by XGBOOST)

The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.5.

In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in

Table 4.5.

According to the results, it can be found that the difference between the test set and the training set is large.

The PR curve of the training set obviously includes the dotted line of the training set PR. The training set AUC is larger than the testing set AUC, but it is not obvious, indicating that the model has a slight over-fitting.

From the results of cross-validation, the conclusions are similar, and the effect of the model in the testing set fluctuates greatly, indicating that the model is not robust in this scenario.

Table 4.5 Result of 10-fold cross validation (by XGBOOST)

	Train					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.922	0.905	0.944	0.924	2.691	0.972
1	0.940	0.938	0.942	0.940	2.065	0.985
2	0.932	0.940	0.922	0.931	2.364	0.986
3	0.936	0.935	0.937	0.936	2.215	0.986
4	0.936	0.938	0.934	0.936	2.215	0.985
5	0.952	0.950	0.955	0.953	1.642	0.985
6	0.945	0.950	0.938	0.944	1.916	0.990
7	0.935	0.930	0.941	0.936	2.240	0.984
8	0.947	0.943	0.951	0.947	1.841	0.980
9	0.938	0.944	0.931	0.938	2.140	0.989
Ave	0.938	0.937	0.939	0.938	2.133	0.984
	Test					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.872	0.854	0.897	0.875	4.428	0.915
1	0.948	0.986	0.909	0.946	1.794	0.969
2	0.955	0.986	0.922	0.953	1.570	0.994
3	0.942	0.905	0.987	0.944	2.019	0.989
4	0.929	0.946	0.909	0.927	2.467	0.975
5	0.870	0.820	0.948	0.880	4.486	0.960
6	0.929	0.902	0.961	0.931	2.467	0.959
7	0.942	0.936	0.948	0.942	2.019	0.991
8	0.851	0.770	1.000	0.870	5.159	0.902
9	0.825	0.791	0.883	0.834	6.056	0.916
Ave	0.906	0.890	0.936	0.910	3.246	0.957

4.1.6 Naive bayes

Moreover, this study constructs a Naive Bayes model to predict credit default by training it with enterprises’ credit data.

The empirical distribution diagram of the degree of confidence, which is predicted by the Naive Bayes model trained by training set, is shown in the top of Figure 4.6. The model outputs results in the training set and testing set of ROC (The Corresponding Receiver

Operating Characteristic) and AUC (Area Under Curve) are shown in the middle part of Figure 4.6.

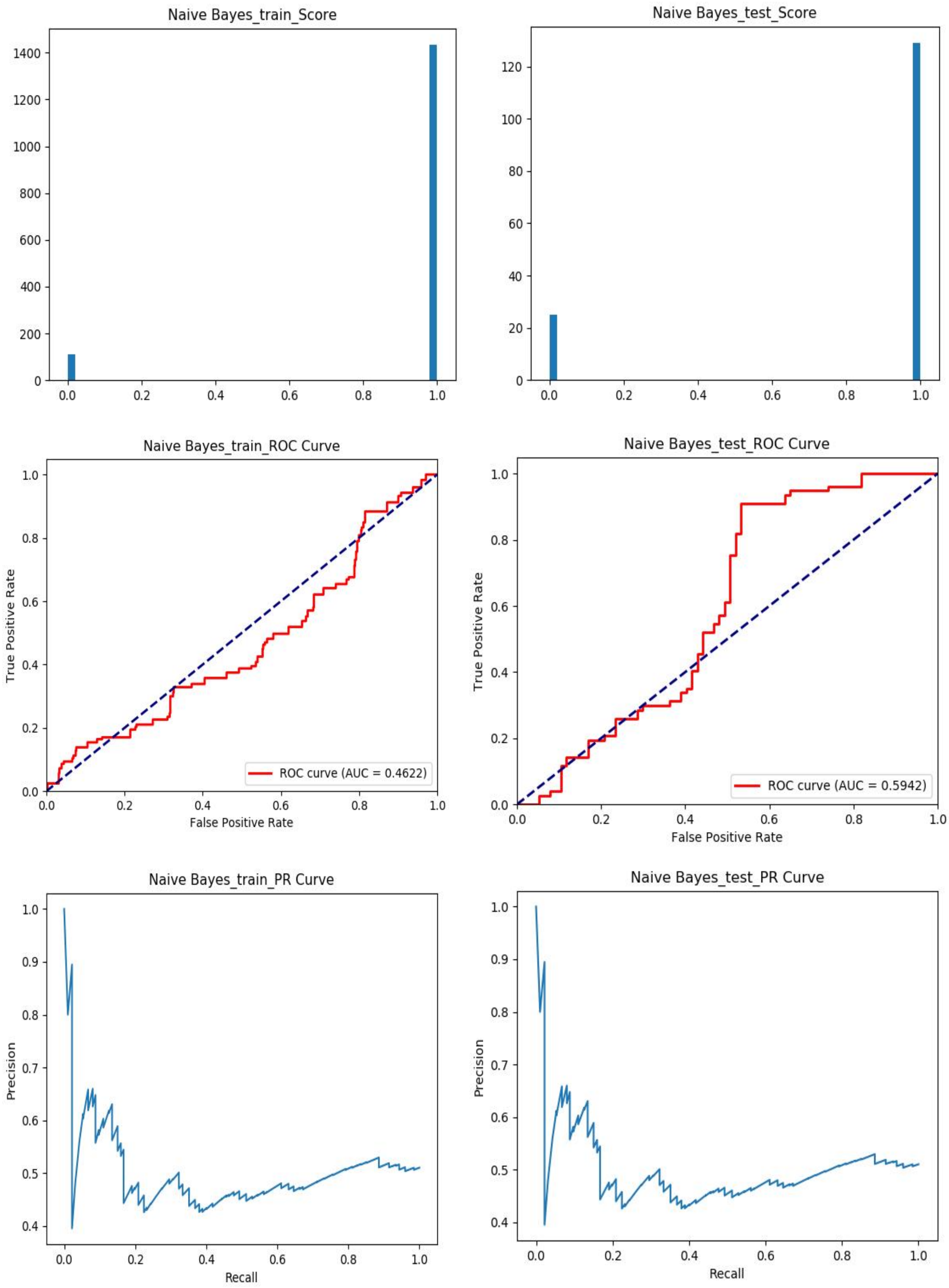


Figure 4.6 Result of 10-fold cross validation (by naive bayes)

The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.6 In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.6.

Table 4.6 Result of 10-fold cross validation (by naive bayes)

			Train			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.517	0.509	0.942	0.661	16.697	0.488
1	0.524	0.513	0.960	0.669	16.424	0.482
2	0.526	0.514	0.961	0.670	16.374	0.484
3	0.516	0.509	0.945	0.661	16.722	0.489
4	0.522	0.512	0.955	0.666	16.523	0.479
5	0.522	0.512	0.957	0.667	16.498	0.479
6	0.513	0.507	0.947	0.660	16.822	0.483
7	0.509	0.505	0.939	0.657	16.971	0.466
8	0.513	0.507	0.950	0.661	16.822	0.473
9	0.503	0.502	0.942	0.655	17.170	0.462
Ave	0.516	0.509	0.950	0.663	16.702	0.479
			Test			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.474	0.487	0.949	0.643	18.155	0.413
1	0.487	0.493	0.961	0.652	17.718	0.452
2	0.474	0.487	0.948	0.643	18.167	0.419
3	0.481	0.490	0.922	0.640	17.943	0.397
4	0.513	0.507	1.000	0.672	16.821	0.456
5	0.506	0.503	0.987	0.667	17.045	0.446
6	0.519	0.511	0.909	0.654	16.597	0.456
7	0.558	0.532	0.974	0.688	15.251	0.534
8	0.539	0.523	0.883	0.657	15.924	0.581
9	0.610	0.566	0.948	0.709	13.457	0.594
Ave	0.516	0.510	0.948	0.663	16.708	0.475

According to the results, it can be found that the PR curve and the ROC curve of the testing set and the training set are both in the vicinity of the diagonal of the range, indicating that the method has almost no recognition effect. From the results of cross-validation, the conclusions are similar.

4.1.7 KNN

This study also has constructed a KNN model to predict credit default by training it with enterprises' credit data.

The empirical distribution diagram of the degree of confidence, which is predicted by the KNN model trained by training set, is shown in the top of Figure 4.7.

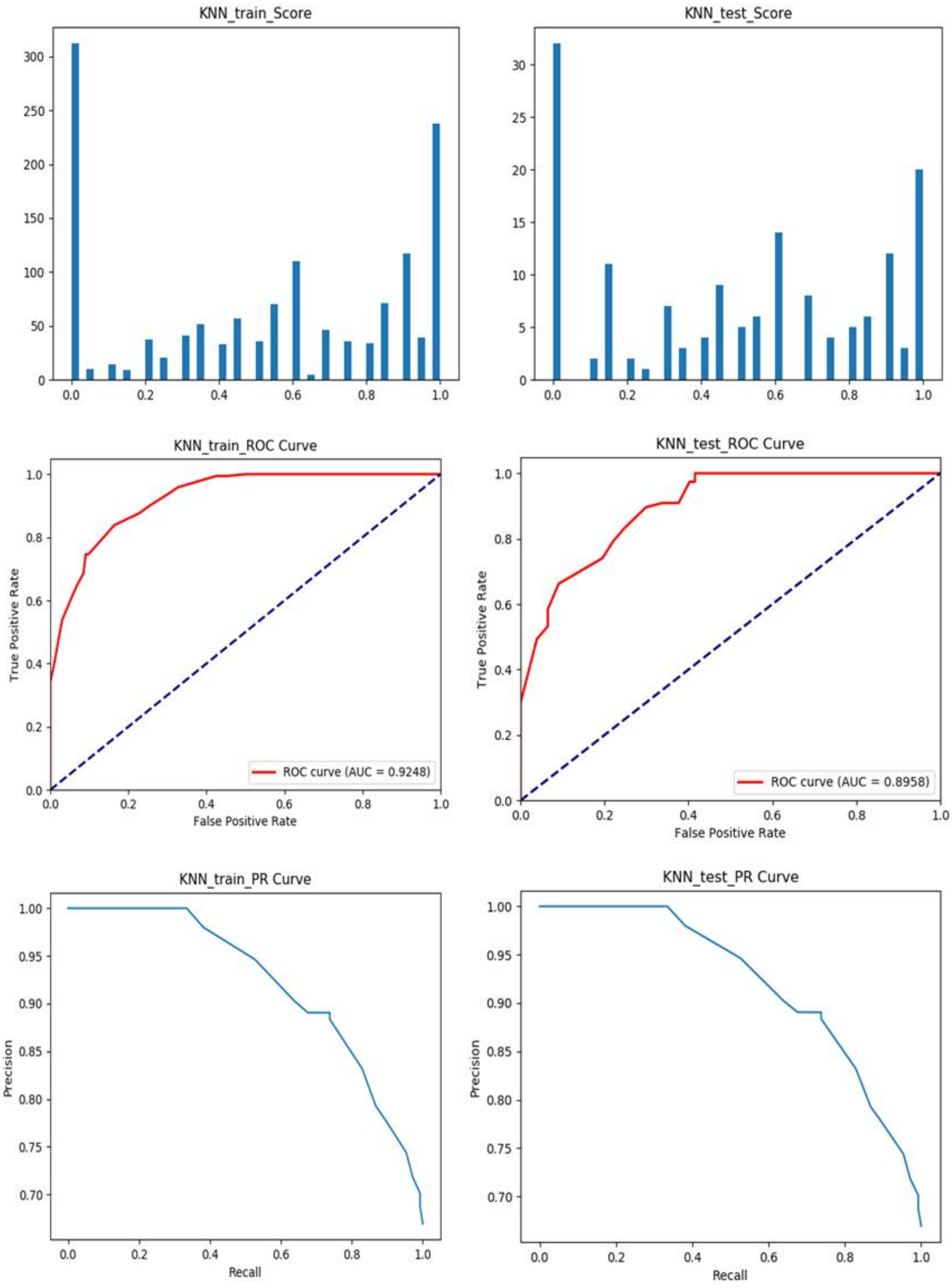


Figure 4.7 Result of 10-fold cross validation (by KNN)

The model outputs results in the training set and testing set of ROC (The Corresponding Receiver Operating Characteristic) and AUC (Area Under Curve) is shown in the middle part of Figure 4.7. The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.7. In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.7.

Table 4.7 Result of 10-fold cross validation (by KNN)

			Train			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.829	0.815	0.851	0.833	5.906	0.927
1	0.836	0.794	0.909	0.848	5.649	0.930
2	0.828	0.788	0.896	0.839	5.947	0.926
3	0.817	0.789	0.865	0.825	6.321	0.921
4	0.834	0.797	0.896	0.844	5.723	0.929
5	0.829	0.799	0.878	0.837	5.922	0.927
6	0.823	0.813	0.840	0.826	6.097	0.927
7	0.826	0.797	0.873	0.834	6.022	0.927
8	0.817	0.806	0.836	0.820	6.321	0.923
9	0.824	0.794	0.876	0.833	6.072	0.925
Ave	0.826	0.799	0.872	0.834	5.998	0.926
			Test			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.776	0.736	0.859	0.793	7.749	0.869
1	0.753	0.697	0.896	0.784	8.523	0.871
2	0.805	0.783	0.844	0.813	6.728	0.915
3	0.838	0.802	0.896	0.847	5.607	0.915
4	0.766	0.730	0.844	0.783	8.074	0.889
5	0.838	0.810	0.883	0.845	5.607	0.917
6	0.792	0.792	0.792	0.792	7.177	0.896
7	0.760	0.738	0.805	0.770	8.298	0.882
8	0.851	0.865	0.831	0.848	5.158	0.935
9	0.786	0.782	0.792	0.787	7.401	0.896
Ave	0.796	0.774	0.844	0.806	7.032	0.899

According to the results, it can be found that the difference between the testing set and the training set is not large, and the PR curve and the distribution of the score and the ROC curve are similar, indicating that there is no significant over-fitting phenomenon.

From the results of cross-validation, there is also no significant over-fitting phenomenon, and the effect of the model in the testing set is less fluctuating, indicating that the model is more stable in this scenario.

4.1.8 SVM

This study constructs an SVM model trained by enterprises' credit data, and uses it to predict credit defaults.

First, all training sets are randomly divided into train set and test set according to the 7:3 ratio. Because the application of the SVM model does not satisfy the required degree of confidence, the results of ROC, PR and distribution are not shown.

In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.8.

Table 4.8 Result of 10-fold cross validation (by SVM)

	Accuracy	Precision	Recall	F1	Loss	AUC
			. Train			
0	0.579	0.543	1.000	0.704	14.529	0.579
1	0.579	0.543	1.000	0.704	14.533	0.579
2	0.577	0.542	1.000	0.703	14.607	0.577
3	0.579	0.543	1.000	0.703	14.557	0.579
4	0.578	0.542	1.000	0.703	14.582	0.578
5	0.578	0.542	1.000	0.703	14.582	0.578
6	0.575	0.540	1.000	0.702	14.682	0.575
7	0.574	0.540	1.000	0.701	14.707	0.574
8	0.571	0.538	1.000	0.700	14.806	0.571
9	0.994	0.999	0.990	0.994	0.199	0.994
Ave	0.618	0.587	0.999	0.732	12.378	0.618
			. Test			
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.500	0.500	1.000	0.667	17.270	0.500
1	0.500	0.500	1.000	0.667	17.270	0.500
2	0.519	0.510	1.000	0.675	16.597	0.519
3	0.506	0.503	1.000	0.670	17.046	0.506
4	0.513	0.507	1.000	0.672	16.821	0.513
5	0.513	0.507	1.000	0.672	16.821	0.513
6	0.526	0.513	1.000	0.678	16.373	0.526
7	0.545	0.524	1.000	0.688	15.700	0.545
8	0.571	0.538	1.000	0.700	14.803	0.571
9	1.000	1.000	1.000	1.000	0.000	1.000
Ave	0.569	0.560	1.000	0.709	14.870	0.569

Since the nonlinear SVM method divides the sample categories according to the hyperplane of the high-dimensional space, the method cannot obtain the confidence that the samples belong to each category, so the method cannot draw the score distribution map, PR curve, and ROC curve.

According to the results, it can be found that the mean value of the AUC of the test set and the training set in the cross-validation does not exceed 0.7. From the accuracy point of view, the average accuracy of the method in the test set is 56%, which is close to the black sample ratio of the sample.

This model does not provide more efficient information than random extraction. In summary, this method has almost no recognition effect.

4.1.9 BP neural network

This study constructs a BP Neural Network model trained by enterprises' credit data, and uses it to predict credit defaults.

The empirical distribution diagram of the degree of confidence, which is predicted by the BP Neural Network model trained by training set, is shown in the top of Figure 4.8.

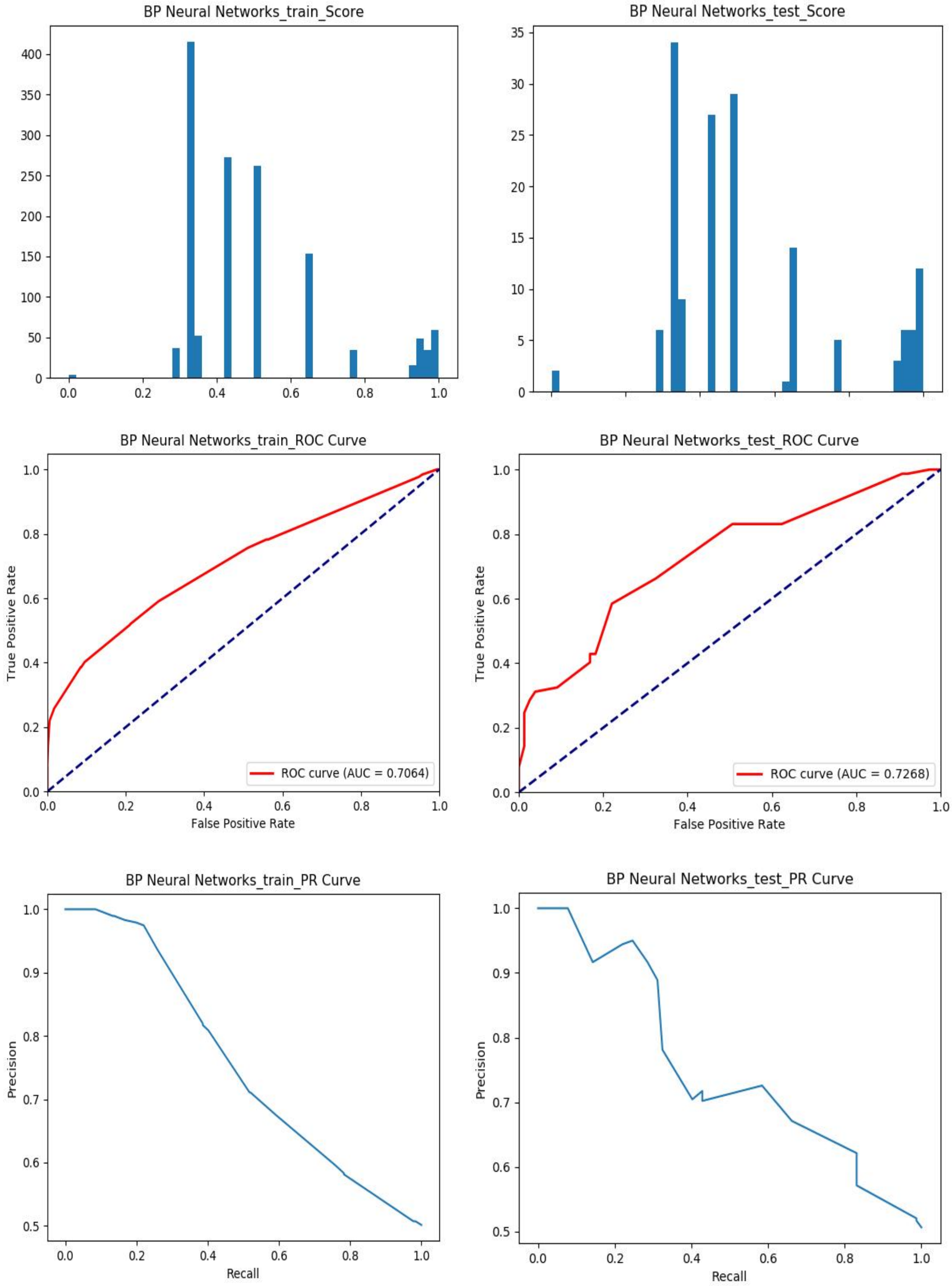


Figure 4.8 Result of 10-fold cross validation (by BP neural networks)

The model outputs results in the training set and testing set of ROC (The Corresponding Receiver Operating Characteristic) and AUC (Area Under Curve) is shown in the middle part of Figure 4.8.

The PR curve of the model results in the training set and testing set is also given in this study, as shown in the bottom of Figure 4.8.

In order to further verify the stability of the model, cross-validation method is used to verify the model at the end of this part. The output results of each training model are shown in Table 4.9.

Table 4.9 Result of 10-fold cross validation (by BP neural network)

	Train					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.662	0.720	0.531	0.611	11.663	0.718
1	0.656	0.709	0.527	0.605	11.895	0.714
2	0.652	0.705	0.523	0.600	12.019	0.706
3	0.660	0.718	0.527	0.608	11.745	0.713
4	0.656	0.713	0.522	0.602	11.895	0.709
5	0.653	0.708	0.522	0.601	11.969	0.709
6	0.653	0.709	0.517	0.598	11.994	0.707
7	0.661	0.686	0.595	0.637	11.696	0.715
8	0.664	0.721	0.533	0.613	11.621	0.720
9	0.653	0.675	0.591	0.630	11.969	0.706
Ave	0.657	0.707	0.539	0.611	11.846	0.712
	Test					
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
0	0.609	0.655	0.462	0.541	13.506	0.655
1	0.669	0.760	0.494	0.598	11.438	0.689
2	0.701	0.804	0.532	0.641	10.317	0.763
3	0.630	0.679	0.494	0.571	12.784	0.703
4	0.669	0.724	0.545	0.622	11.438	0.743
5	0.688	0.764	0.545	0.636	10.765	0.741
6	0.688	0.738	0.584	0.652	10.765	0.737
7	0.604	0.600	0.623	0.611	13.681	0.679
8	0.597	0.642	0.442	0.523	13.905	0.628
9	0.669	0.671	0.662	0.667	11.438	0.727
Ave	0.652	0.704	0.538	0.606	12.004	0.706

According to the results, it can be found that the mean value of the AUC of the test set and the training set in the cross-validation is approximately equal to 0.7.

From the accuracy point of view, the average accuracy of the method in the test set is 65%, which is slightly higher than the black sample of the sample. The ratio indicates that the model provides only a small amount of valid information compared to random extraction, which is much lower than most of the above methods. In summary, the method has a poor recognition effect.

4.2 Synthesis of the results

Firstly, we will review the inference of past literature using machine learning models to evaluate SME credit. Secondly, we will compare its implications with our conclusions.

T. S. Lee et al.(2002) proposed a hybrid credit scoring model that combines neural network and LDA (Linear Discriminant Analysis). Compared with models that use LDA, LRA (Logistic RegressionAnalysis) or ANN (Artificial Neural Network) alone, the performance of this model is more successful.

Abdou et al.(2008) found in a study using Egyptian personal loan data that ANN is more successful than LDA, LRA and Probit analysis. Angelini et al.(2008) used ANN to achieve an average error rate of 7% on a credit dataset composed of small and medium-sized enterprises obtained by an Italian bank.

Huang et al.(2004) used genetic algorithms to transfer three rejected credit applications to the conditional acceptance group, and found that the ANN model is more successful than the LDA and CART models.

T.S. Lee et al.(2006) found that models developed using CART and MARS (Multivariate Adaptive Regression Splines) on credit card data are more successful than models using LDA, LRA, and ANN.

T.S. Lee and Chen(2005) compared the performance of LDA, LRA, ANN, MARS and MARS-ANN models on the mortgage loan dataset of a bank in Taiwan. The ANN model that uses the more important variables discovered by MARS obtains the best performance.

Chuang and Lin(2009) obtained 76%, 76.5%, 77.5%, 79.5% and 82.5% of German credit data prediction performance from LDA, LRA, CART, ANN and MARS-ANN models, respectively.

In the last part of the study, when the data transferred to the bad credit group was re-evaluated based on Case Based Reasoning (CBR), the accuracy of the model reached 86%.

Tsai et al.(2009) found that data envelopment analysis-LDA and ANN models are more successful than LDA and LRA on Taiwan's consumer credit data.

In recent years, in order to improve the performance of credit scoring models, people have proposed integrated classifiers. The main idea of ensemble classifiers is to combine multiple classifiers into a multi-classifier(Nanni & Lumini, 2006).

West et al.(2005) determined that the ensemble classifier-neural network model reduces the error rate of a single classifier by 3% or 5%. Yu et al.(2008) found that in a single classifier, ANN and SVM are more successful than LRA, and the integrated classifier-ANN has the best performance.

Similarly, in the study of Nanni and Lumini(2006), we determined that ANN is the best in a single classifier, but usually the best performance is obtained through the random subspace ensemble of the Levenberg Marquardt neural network model classifier.

In the study of Tsai et al.(2009), the ensemble classifier-ANN performed well in only one of the three datasets. Hsieh and Hung(2010) used cluster analysis to classify German credit data into three categories: good, bad and marginal, and developed an integrated classifier credit scoring model.

Finlay(2011) compared the performance of multiple classifiers and found that Error trim Boosting outperforms all other multiple classifiers on UK credit data. In this study almost all of these methods are covered instead of LDA because it is not considered as a machine learning method. Although the dataset, variable selection, hyperparameter tuning, algorithm details have inconsistency, some consistent conclusion is obtained.

Firstly, almost all the machine learning methods are more effective than the linear methods such as LDA and LRA (in this study we named it Logistic Regression). Secondly, the Tree model is good for evaluation of loan credit risk. Thirdly, Neural Network (also named ANN) may work well for evaluation. Lastly, ensemble learning can improve the accuracy obviously.

In fact, three of the best performing methods in this study are ensemble learning methods.

Table 4.10 shows the results of all the models in the test set and training set, including the model before and after the irrelevant variables are removed.

Table 4.10 Comparison of different models' results

	The Effect of Train Set					
	Accuracy	**precision**	**recall**	**F1**	**loss**	**AUC**
Logistic	0.866	0.878	0.849	0.863	4.637	0.943
Logistic-After Selecting	0.844	0.853	0.844	0.845	5.378	0.920
Decision Tree	0.871	0.893	0.845	0.868	4.44	0.931
Random Forest	0.82	0.906	0.716	0.799	6.207	0.888
Random Forest-After Selecting	0.806	0.865	0.734	0.792	6.698	0.871
Naïve Bayes	0.516	0.509	0.95	0.663	16.702	0.479
KNN	0.826	0.799	0.872	0.834	5.998	0.926
SVM	0.618	0.587	0.999	0.732	12.378	0.618
BP Neural Network	0.657	0.707	0.539	0.611	11.846	0.712
ADABOOST	0.851	0.909	0.780	0.839	5.149	0.950
	The Effect of Train Set					
	Accuracy	**precision**	**recall**	**F1**	**loss**	**AUC**
ADABOOST-After Selecting	0.838	0.902	0.764	0.825	5.579	0.926
XGBOOST	0.938	0.937	0.939	0.938	2.133	0.984
XGBOOST-After Selecting	0.906	0.890	0.936	0.910	3.246	0.957
Logistic	0.842	0.850	0.840	0.842	5.467	0.919
Logistic-After Selecting	0.848	0.861	0.842	0.848	5.243	0.918
Decision Tree	0.84	0.849	0.834	0.84	5.509	0.909
Random Forest	0.806	0.865	0.733	0.792	6.698	0.871
Random Forest-After Selecting	0.830	0.901	0.746	0.815	5.869	0.902
Naive Bayes	0.516	0.51	0.948	0.663	16.708	0.475
KNN	0.796	0.774	0.844	0.806	7.032	0.899
SVM	0.569	0.56	1	0.709	14.87	0.569
BP Neural Network	0.652	0.704	0.538	0.606	12.004	0.706
ADABOOST	0.831	0.873	0.778	0.822	5.826	0.935
ADABOOST-After Selecting	0.939	0.911	0.975	0.941	2.107	0.985

XGBOOST	0.906	0.89	0.936	0.91	3.246	0.957
XGBOOST-After Selecting	0.934	0.913	0.961	0.936	2.285	0.987

It can be seen that, from the overall effect, the ADABOOST and XGBOOST models after the variables are removed are much better than other models in terms of accuracy, precision, recall, F1, and AUC.

At the same time, although the XGBOOST and ADABOOST methods are very similar in the test set, comparing the difference between the two models in the training set and the test set, it can be seen that the training set effect of the XGBOOST model is closer to the test set effect, indicating that the XGBOOST model is over-extended. The problem is lighter.

Therefore, the XGBOOST model after the variables are removed is the best model in this study.However, this does not mean that XGBOOST and ADABOOST are the optimal methods for any commercial bank. The reason is that XGBOOST needs to pre order the characteristics of nodes before iteration, and traverse to select the optimal segmentation point. When a commercial bank has a large number of samples, this method is too time-consuming; XGBOOST uses level wise to generate decision trees, and splits leaves at the same decision level for multithreaded optimization. Although it is not easy to cause over fitting problems, many leaf nodes have low splitting gains, which increases the error; However, ADABOOST is often difficult to set the number of iterations, that is, weak classifiers, and the problem of category imbalance easily leads to the decline of classification accuracy. The training cost of this method is too high and time-consuming, and the best segmentation point of the current classifier needs to be re-selected each time. Therefore, when selecting the machine learning model for credit risk assessment, commercial banks need to make specific analysis according to the specific problems of bank needs and data sets.

Table 4.11 Comparison of different researches

	The Effect of Train Set		
	Accuracy	**AUC**	**Source**
SVM	0.854	0.935	Zhou&Wang(2015)
SVM	0.834	-	Chen(2012)
BP Neural Network	0.795	-	Chen(2012)
ANN	0.758	-	Wang&Yao(2018)
SVM	0.808	-	Wang&Yao(2018)
ADABOOST	0.799	-	Wang&Yao(2018)
SVM	0.752	-	Xia(2015)
BP Neural Network	0.789	-	Tan(2009)

At the same time, this research compares the application results of the model with similar studies using Chinese enterprise data in the references, and the specific results are shown in Table 4.11. Most studies only return the accuracy of model fitting, so this study compares the accuracy. In contrast, the estimation accuracy of this research using BP neural network and SVM is significantly lower than that of existing research, but the estimation result of ADABOOST is better than that of existing research.

4.3 Feature selection

Since the data used for model training are limited rather than unlimited, the parameter estimation accuracy of the model is affected by the number of samples participating in the training and the sample feature dimension.

In general, the addition of useless features affects the performance of model predictions when the number of samples is small(Alpaydin, 2020; Domingos, 2012). Therefore, the characteristics of each model feature in the model are selected and the model is re-trained. Since some models cannot calculate the importance of features to the model, this section only reports models that allow it.

The mechanism of importance measure is completely different for different models. Most machine learning models are no-linear even unknowable, just as several neural network models, so that people cannot distinguish the role of features in the model(Alpaydin, 2020)

In this study, SVM, Naive Bayes, KNN and BP Neural Network are difficult to measure feature importance. SVM maps low-dimensional space to unobservable higher-dimensional space. Naive Bayes is dependent on the Bayes network where the relation between features is conditional probability. The accuracy of KNN classification depends largely on the Mahalanobis distance between samples. BP Neural Network classifies samples by simulating human neural system.

The common point of these models is that the calculation of their eigenvalues depends on complex mathematical functions and can not be obtained directly. In contrast, logistic model and decision tree model are simplified models. Logistic model belongs to linear model, and eigenvalues can be obtained directly. The decision tree model directly models the eigenvalues through a simple binary tree. Therefore, these two models are easy to explain. Random forest,ADABOOST and XGBBOOST are models that combine multiple decision trees, that is, multiple single mechanisms constitute the review mechanism. If the composite mechanism is linear, it is also easy to explain. Therefore, we only reported the feature selection results of logistic, decision tree, random forest,ADABOOST and XGBBOOST models to obtain the feature importance.

4.3.1 Logistic regression

The Logistic model can be written as equation :

$$\mathbf{P} = \frac{\mathbf{1}}{\mathbf{1} + e^{-(\beta_1 x_1 + \ldots + \beta_n x_n)}} \tag{4.1}$$

So, the gradient of it is equation :

$$\Delta = \frac{e^{-(\beta_1 x_1 + \ldots + \beta_n x_n)}}{\left[\mathbf{1} + e^{-(\beta_1 x_1 + \ldots + \beta_n x_n)}\right]^2} \times \{\beta_1, \ldots, \beta_n\} \tag{4.2}$$

It means that the more β_i is, the more effect on P is, and it is equivalent to the meaning of importance. So, β_i can represent the feature importance of feature(McCulloch & Rossi, 1994). Because we use L1 and L2-norm , parameters of unimportance feature will contraction to zero. So we just select the feature whose parameters are not zero.

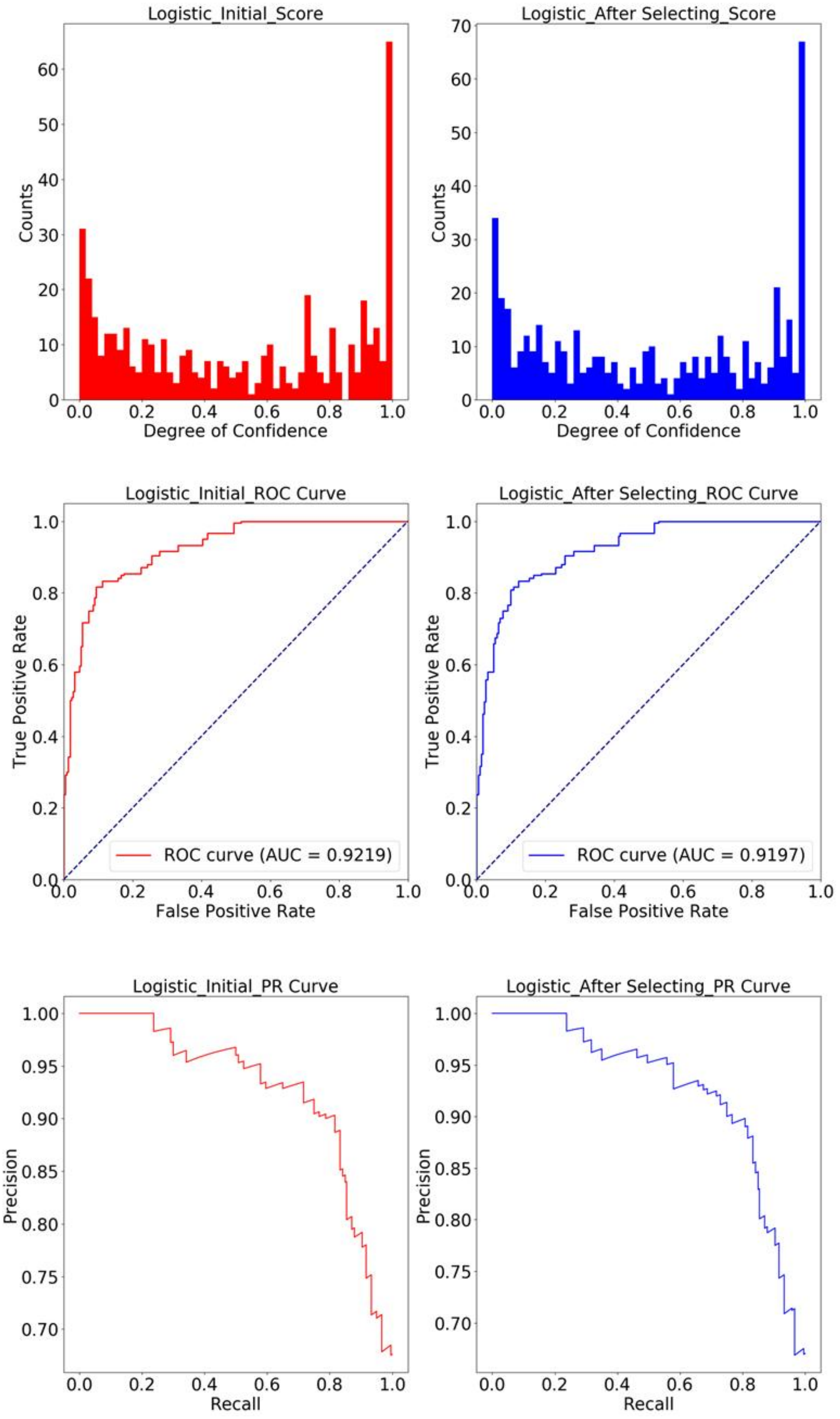
In the Logistic model, unimportant variables are proposed using the L1 penalty term. We can see the comparison of the effects of the model before and after the unimportant variables in Figure 4.9 Panel (A).

4.3.2 Randomforest

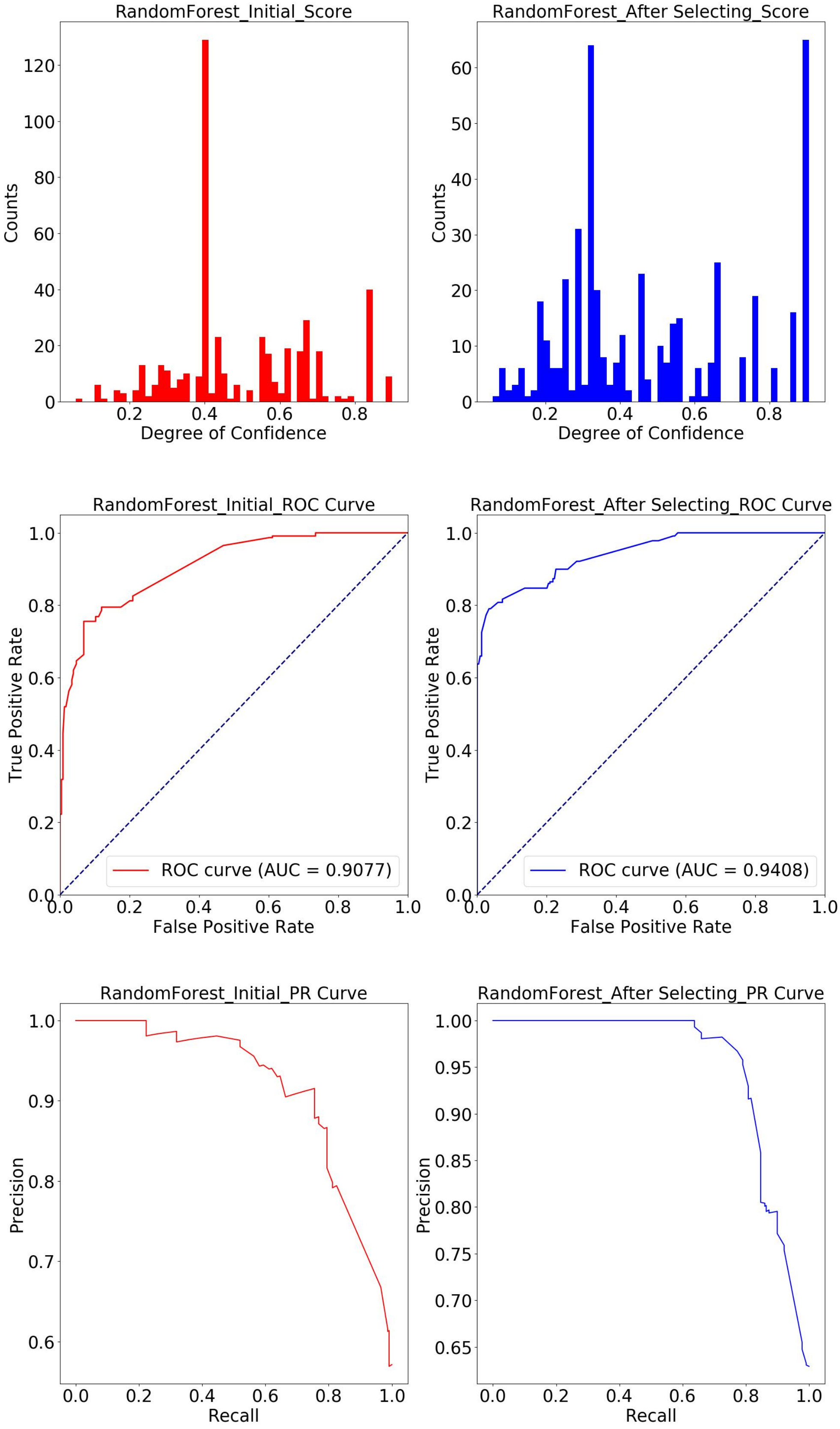
Feature importance of Random Forest is based on a simple idea: the change of absolute value of the error before and after adding noise to some feature reflects the importance of the feature(Stroblet al., 2007). We select the feature whose importance index is bigger that 5%.

We can see Figure 4.9 Panel (B) for a comparison of the effects of the model before and after the unimportant variables is removed. Cross-validation is used to verify the effectiveness of the model in the validation set before and after culling variables in Table 4.11 Panel (B).

After eliminating the unimportant variables, the accuracy, precision, recall, F1, and AUC indicators are higher than before the culling, and the loss index becomes lower, indicating that the culling variable does have some optimization effect on the model.

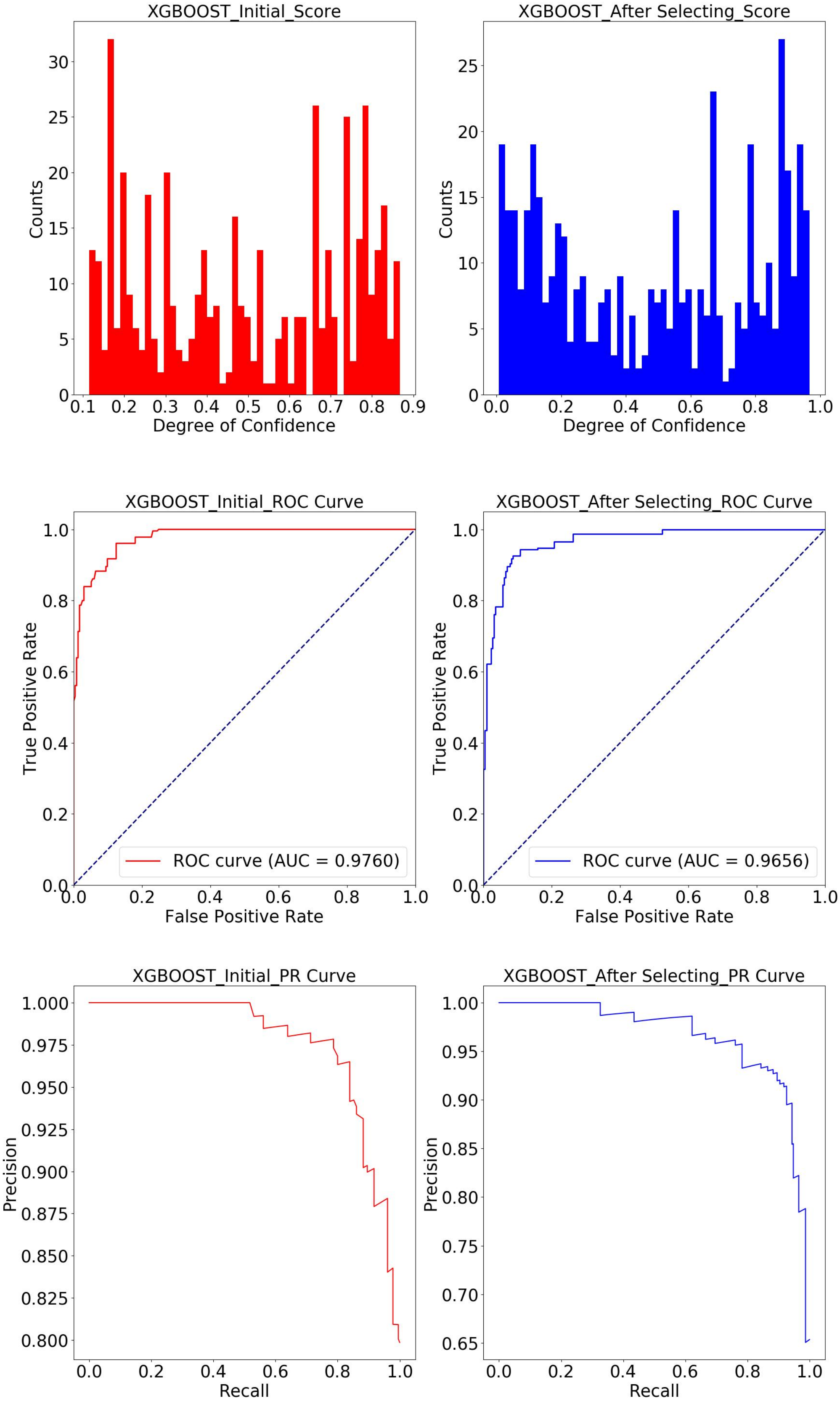


Panel (A) Logistic regression



Panel (B)Random forest

Panel (C)ADABOOST



Panel (D) XGBOOST

Figure 4.9 Performance testing set before and after feature selection

Cross-validation is used to verify the effectiveness of the model in the validation set before and after culling variables in Table 4.12 Panel (A). It can be seen that the Logistic model has no significant effect before and after the variables are eliminated.

Table 4.12 Result of feature selection

Panel (A) Result of 10-fold cross validation (by Logistic Regression)

			Initial			
	accuracy	**precision**	**Recall**	**F1**	**loss**	**AUC**
0	0.891	1.000	0.782	0.878	3.764	0.935
1	0.779	0.795	0.753	0.773	7.626	0.896
2	0.890	0.929	0.844	0.884	3.813	0.962
3	0.818	0.795	0.857	0.825	6.280	0.915
4	0.877	0.914	0.831	0.871	4.261	0.938
5	0.779	0.747	0.844	0.793	7.626	0.884
6	0.870	0.852	0.896	0.873	4.486	0.945
7	0.903	0.943	0.857	0.898	3.364	0.963
8	0.844	0.791	0.935	0.857	5.383	0.898
9	0.792	0.765	0.844	0.802	7.177	0.862
Ave	0.844	0.853	0.844	0.845	5.378	0.920

			After selecting			
	accuracy	**precision**	**Recall**	**F1**	**loss**	**AUC**
0	0.891	1.000	0.782	0.878	3.764	0.933
1	0.779	0.795	0.753	0.773	7.626	0.892
2	0.896	0.942	0.844	0.890	3.588	0.961
3	0.825	0.805	0.857	0.830	6.056	0.911
4	0.877	0.914	0.831	0.871	4.261	0.937
5	0.786	0.756	0.844	0.798	7.401	0.882
6	0.857	0.848	0.870	0.859	4.934	0.944
7	0.909	0.957	0.857	0.904	2.340	0.962
8	0.857	0.809	0.935	0.867	4.934	0.906
9	0.805	0.783	0.844	0.813	6.728	0.854
Ave	0.848	0.861	0.842	0.848	5.243	0.918

Panel (B) Result of 10-fold cross validation (by random forest)

			Initial			
	accuracy	**precision**	**Recall**	**F1**	**loss**	**AUC**
0	0.827	0.84	0.808	0.824	5.978	0.89
1	0.812	0.853	0.753	0.8	6.504	0.87
2	0.87	0.938	0.792	0.859	4.486	0.903
3	0.792	0.846	0.714	0.775	7.177	0.878
4	0.844	0.949	0.727	0.824	5.383	0.878
5	0.812	0.914	0.688	0.785	6.504	0.878
6	0.799	0.859	0.714	0.78	6.953	0.881
7	0.844	0.965	0.714	0.821	5.383	0.903
8	0.747	0.757	0.727	0.742	8.747	0.806
9	0.714	0.726	0.688	0.707	9.868	0.823
Ave	0.806	0.865	0.733	0.792	6.698	0.871

			After selecting			
	accuracy	**precision**	**Recall**	**F1**	**loss**	**AUC**
0	0.840	0.844	0.833	0.839	5.535	0.900
1	0.870	0.983	0.753	0.853	4.486	0.880
2	0.890	1.000	0.779	0.876	3.813	0.895
3	0.831	0.918	0.727	0.812	5.831	0.912
4	0.857	0.982	0.727	0.836	4.934	0.918
5	0.818	0.855	0.766	0.808	6.280	0.890

After selecting

	accuracy	precision	Recall	F1	loss	AUC
6	0.825	0.917	0.714	0.803	6.056	0.925
7	0.838	0.894	0.766	0.825	5.607	0.947
8	0.786	0.814	0.740	0.776	7.401	0.909
9	0.747	0.806	0.649	0.719	8.747	0.848
Ave	0.830	0.901	0.746	0.815	5.869	0.902

Panel (C) Result of 10-fold cross validation (by ADABOOST)

Initial

	accuracy	precision	Recall	F1	loss	AUC
4	0.857	0.982	0.727	0.836	4.934	0.947
5	0.805	0.862	0.727	0.789	6.728	0.883
6	0.825	0.879	0.753	0.811	6.056	0.943
7	0.831	0.932	0.714	0.809	5.831	0.967
8	0.792	0.792	0.792	0.792	7.177	0.911
9	0.799	0.838	0.740	0.786	6.953	0.895
Ave	0.838	0.902	0.764	0.825	5.579	0.926

After selecting

	accuracy	precision	Recall	F1	loss	AUC
0	0.962	0.939	0.987	0.963	1.328	0.994
1	0.942	0.905	0.987	0.944	2.019	0.993
2	0.916	0.856	1.000	0.922	2.916	0.990
3	0.955	0.917	1.000	0.957	1.570	0.996
4	0.955	0.973	0.935	0.954	1.570	0.995
5	0.948	0.906	1.000	0.951	1.794	0.997
6	0.922	0.865	1.000	0.928	2.691	0.979
7	0.942	0.936	0.948	0.942	2.019	0.988
8	0.968	0.939	1.000	0.969	1.121	0.994
9	0.883	0.873	0.896	0.885	4.037	0.927
Ave	0.939	0.911	0.975	0.941	2.107	0.985

Panel (D) Result of 10-fold cross validation (by XGBOOST)

Initial

	accuracy	Precision	Recall	F1	loss	AUC
0	0.872	0.854	0.897	0.875	4.428	0.915
1	0.948	0.986	0.909	0.946	1.794	0.969
2	0.955	0.986	0.922	0.953	1.570	0.994
3	0.942	0.905	0.987	0.944	2.019	0.989
4	0.929	0.946	0.909	0.927	2.467	0.975
5	0.870	0.820	0.948	0.880	4.486	0.960
6	0.929	0.902	0.961	0.931	2.467	0.959
7	0.942	0.936	0.948	0.942	2.019	0.991
8	0.851	0.770	1.000	0.870	5.159	0.902
9	0.825	0.791	0.883	0.834	6.056	0.916
Ave	0.906	0.890	0.936	0.910	3.246	0.957

After selecting

	accuracy	Precision	Recall	F1	Loss	AUC
0	0.936	0.936	0.936	0.936	2.214	0.985
1	0.935	0.904	0.974	0.938	2.243	0.990
2	0.955	0.938	0.974	0.955	1.570	0.995
3	0.961	0.928	1.000	0.963	1.346	0.998
4	0.942	0.947	0.935	0.941	2.019	0.992
5	0.974	0.962	0.987	0.974	0.897	0.995
6	0.909	0.871	0.961	0.914	2.340	0.976

7	0.890	0.849	0.948	0.896	3.813	0.982
8	0.929	0.875	1.000	0.933	2.467	0.988
9	0.909	0.920	0.896	0.908	2.340	0.967
Ave	0.934	0.913	0.961	0.936	2.285	0.987

4.3.3ADABOOST

ADABOOST is a higher-level decision tree of forest type, and its feature importance is similar to that of decision tree. The purity of dataset D can be measured by Gini coefficient as the following equation :

$$\boldsymbol{Gini}\left(\boldsymbol{D}\right) = 1 - \Sigma\, \boldsymbol{p}_i^2 \tag{4.3}$$

Reflects the level of category labels are inconsistent of two samples which are randomly selected from dataset D. Then, the Gini index of feature a is defined as equation :

$$\boldsymbol{Gini}_{index}\left(\boldsymbol{D},\boldsymbol{a}\right) = \Sigma \frac{\left|\boldsymbol{D}^j\right|}{\left|\boldsymbol{D}\right|} \boldsymbol{Gini}\left(\boldsymbol{D}^j\right) \tag{4.4}$$

The smaller the Gini index, the greater the ‘purity boost’ which is equivalent to feature importance achieved(Hastie et al., 2009). Then, ADABOOST is a linear function of results of several decision tree, and therefore the feature importance of ADABOOST is a linear function of feature importance of decision trees.

We select the feature whose importance index is bigger that 5%. We can see Figure 4.9 Panel (C) for a comparison of the effects of the model before and after the unimportant variables is removed. Cross-validation is used to verify the effectiveness of the model in the validation set before and after culling variables in Table 4.11 Panel (C).

After eliminating the unimportant variables, the accuracy, precision, recall, F1, and AUC indicators are significantly improved before the elimination, and the loss index becomes lower, indicating that the unimportant variables have a greater impact on the model.

4.3.4 XGBOOST

Compared withADABOOST and random forest, XGBOOST is more complex and it is difficult to accurately measure the importance of features. Therefore, it only considers the number of feature segmentation nodes. The more segmentation times, the more important the feature is (J. H. Chenet.al, 2005). In this part, only features with an importance index greater than 2 are selected.

In Figure 4.9 Panel (D) the comparison of the effects of the model before and after the unimportant variables is removed can be seen. Cross-validation is used to verify the effectiveness of the model in the validation set before and after culling variables in Table 4.12 Panel (D).

After eliminating the unimportant variables, the model has a slight increase in accuracy, precision, recall, F1, and AUC indicators, and the loss index becomes lower, indicating that the unimportant variables have a certain impact on the model.

4.4 Oversampling andundersampling

One of the mostimportant problems with the datasets used in this study is unbalanced class. Since this study uses multiple algorithms to test the same dataset, the improvement based on the algorithm level is too complex, so this section focuses on the improvement based on the sample level.

In this section, there are five methods which used to deal with unbalanced class such as: Random Oversampling, SMOTE, Borlderline SMOTE, SVM SMOTE, Random Oversampling and SMOTE Tomek Linksmethod.

In this study, the samples of subsequent defaults are recorded as black samples, and the samples of non default loans are recorded as white samples. All sample data are constructed into data sets, which are divided into training sets and test sets according to 7:3. The report of the result of the training set and the test set is as following (Table 4.13).

Table 4.13 Different method of oversampling andunder-sampling

Panel (A) Logistic

	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
	Train					
SMOTE	0.85	0.85	0.84	0.84	5.31	0.93
BorderlineSMOTE	0.87	0.88	0.85	0.87	4.56	0.95
SVMSMOTE	0.89	0.92	0.86	0.89	3.70	0.96
UnderSampling	0.87	0.90	0.84	0.87	4.33	0.96
SMOTE Tomek Links	0.90	0.91	0.90	0.90	3.39	0.97
	Test					
SMOTE	0.82	0.82	0.83	0.82	6.21	0.91
BorderlineSMOTE	0.85	0.86	0.85	0.85	5.20	0.93
SVMSMOTE	0.87	0.90	0.84	0.86	4.55	0.96
UnderSampling	0.79	0.86	0.72	0.77	7.12	0.89
SMOTE Tomek Links	0.88	0.88	0.88	0.88	4.25	0.95

Panel (B) Random forest

	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
	Train					
SMOTE	0.81	0.84	0.78	0.81	6.50	0.89
Borderline SMOTE	0.87	0.89	0.84	0.87	4.52	0.94
SVM SMOTE	0.87	0.93	0.80	0.86	4.58	0.94
UnderSampling	0.95	0.94	0.97	0.95	1.60	0.99
	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**

	Train					
SMOTE Tomek Links	0.88	0.91	0.85	0.88	4.12	0.95
	Test					
SMOTE	0.77	0.79	0.76	0.77	7.88	0.84
Borderline SMOTE	0.86	0.87	0.85	0.86	4.91	0.93
SVM SMOTE	0.82	0.89	0.74	0.78	6.27	0.94
UnderSampling	0.88	0.89	0.87	0.87	4.10	0.94
SMOTE Tomek Links	0.84	0.87	0.83	0.84	5.48	0.92

Panel (C)ADABOOST

	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
	Train					
SMOTE	0.86	0.92	0.79	0.85	4.82	0.95
Borderline SMOTE	0.94	0.98	0.90	0.94	2.12	0.98
SVM SMOTE	0.94	0.97	0.91	0.94	2.06	0.99
UnderSampling	0.50	0.00	0.00	0.00	17.27	0.50
SMOTE Tomek Links	0.91	0.96	0.86	0.91	3.09	0.97
	Test					
SMOTE	0.83	0.87	0.78	0.82	5.85	0.92
Borderline SMOTE	0.92	0.95	0.90	0.92	2.68	0.97
SVM SMOTE	0.92	0.96	0.88	0.91	2.75	0.98
UnderSampling	0.50	0.00	0.00	0.00	17.27	0.50
SMOTE Tomek Links	0.89	0.94	0.86	0.89	3.63	0.97

Panel (D) XGABOOST

	Accuracy	**Precision**	**Recall**	**F1**	**Loss**	**AUC**
	Train					
SMOTE	0.95	0.94	0.96	0.95	1.85	0.99
Borderline SMOTE	0.96	0.98	0.94	0.96	1.47	0.99
SVM SMOTE	0.95	0.97	0.93	0.95	1.73	0.99
UnderSampling	0.91	0.89	0.93	0.91	3.22	0.97
SMOTE Tomek Links	0.95	0.98	0.92	0.95	1.86	0.98
	Test					
SMOTE	0.91	0.88	0.95	0.91	3.25	0.95
Borderline SMOTE	0.91	0.95	0.88	0.88	3.10	0.97
SVM SMOTE	0.89	0.92	0.86	0.87	3.86	0.97
UnderSampling	0.83	0.83	0.81	0.81	5.98	0.87
SMOTE Tomek Links	0.92	0.95	0.90	0.92	2.66	0.97

We can see that the three modified version of SMOTE method which are SMOTE Tomek Links, Borderline SMOTE and SMOTE Tomek Links are often better than Random Over-sampling, SMOTE method and Random Under-sampling (the result can be found at Section 4.1).

4.5 Cooperative training

Another important problem with the datasets used in this study is that some labels may be incorrect.

So we also test the typical method used to deal with it which is named cooperative training. Cooperative training is similar to ensemble learning which used two or more model to identify the sample class.

The Tri-training method denoted by Zhou & Li(2005) is tested which has been introduced in Section 3.4.2. The three basic models are ADABOOST, XGBOOST and Logistic. The hyperparameter is the same as in Section 4.1 allowing for easy comparison.

The uncreditable sample is here named unlabeled data, the true labeled sample is named labeled data. The unmatured debts which may possibly default in the future are defined as unlabeled data and matured debts as labeled data. The Tri-training results are reported as following (Table 4.14).

We can see that training only with labeled data is more effective than all data. However, Tri-training with unlabeled data and labeled data is better than training only with labeled data.

It states that unmatured debts is a noise for machine learning model, but Tri-training or cooperative training method can utilize the useful information in the noise to make model more effective. All in all, Tri-training method can improve the effect of model.

4.6 Important feature attributes in the evaluation

Among all the methods in this study, the Logistic, Decision Tree, Random Forest, ADABOOST, and XGBOOST methods can sort the importance of features. The top seven feature attributes that have significant effects on the prediction result of these models have been highlighted in the Table 4.14.

Table 4.14 Tri-Training results

	ROC	**Accuracy**	**F1**	**Loss**	**Precision**	**Recall**
			All Data			
ADABOOST	0.9922	0.9438	0.9427	1.9396	0.9511	0.9345
XGBOOST	0.9881	0.9568	0.9582	1.4920	0.9197	1.0000
Random Forest	0.8627	0.8035	0.7764	6.7884	0.8876	0.6900
			Only Labeled Data			
ADABOOST	0.9966	0.9617	0.9628	1.3221	0.9673	0.9583
XGBOOST	0.9953	0.9617	0.9643	1.3221	0.9310	1.0000
Random Forest	0.8692	0.8230	0.8204	6.1146	0.8622	0.7824
			Tri-training			
ADABOOST	0.9917	0.9698	0.9601	1.3352	0.9756	0.9259
XGBOOST	0.9972	0.9809	0.9818	0.6610	0.9643	1.0000
Random Forest	0.9013	0.8358	0.8153	6.3625	0.8458	0.7870

As is shown in Table 4.14, attributes Life, Maturity, BJJE, LLFDZ, Registered Capital, FFJE ZXNLL all have significant effects on the final result of prediction.

Table 4.14 Importance of feature and variable mapping

Panel(A)

	Logistic	**Decision Tree**	**Random Forest**	**ADABOOST**	**XGBOOST**	**Ave**
Life	0.0024	1.0000	0.5369	1.0000	1.0000	0.7079
	Logistic	**Decision Tree**	**Random Forest**	**ADABOOST**	**XGBOOST**	**Ave**
Maturity	0.0007	0.1231	1.0000	0.8342	0.8342	0.5585
BJJE	0.0000	0.0397	0.6305	0.3992	0.3992	0.2937
LLFDZ	0.0217	0.0000	0.5288	0.4413	0.4413	0.2866
Registered Capital	0.0000	0.2912	0.1504	0.3634	0.3634	0.2337
FFJE	0.5345	0.0000	0.1232	0.1949	0.1949	0.2095
ZXNLL	0.3454	0.0000	0.1407	0.2598	0.2598	0.2012

Panel(B)

Variable Name	**Meaning of the variable**
Life	Time since the Loan
Maturity	Time until the Maturity of the Loan
BJJE	Remained Amount of the Capital
LLFDZ	Fluctuating Interest Rate
Registered Capital	Registered Capital
FFJE	Amount of the Loan
ZXNLL	Actual Annual Interest Rate

The values in the table indicate the characteristic influence of the feature normalization. In the same method, the feature with the highest influence is normalized to 1, and vice versa.

Ave represents the average of the influences of the five model features and represents the overall feature importance. The table is ranked according to the level of Ave indicators. The higher the indicator, the higher the influence of the indicator on loan default.

To be specific, the longer the time is since the enterprise got the loan, the more likely is the case that the enterprise isn’t short for money for only a short time but has problems with their capital chain in their management, indicating a higher probability of default. A larger amount of money need to cover the capital means more pressure faced by the enterprise, which will also increase the credit risk.

Enterprises with higher fluctuating interest rates always have more cost for loans, then more cash pressure and finally higher risks of default. Similarly, if the maturity date for an enterprise is drawing near, the enterprise will not only need to pay the capital and interest at the current period but also need to pay back the capital, interest and penalty of unpaid loans. Thus the probability of default of such an enterprise will be relatively higher.

The actual annual interest rate will also affect the credit risk. If the actual interest rate is high, the cost for the enterprise to get a loan will also be high, which will also cause the credit risk to be high. Finally, the higher the amount of loan is, the higher pressure of the enterprise is, making a higher probability of default.

Chapter 5: Conclusion

5.1Conclusions

The indicator system of Chinese SMEs credit assessment measures is gradually shifting from financial indicators to non-financial indicators, and more attention has been paid to new indicators such as growth, innovation, industry environment, as well as the overall quality of enterprises and managers.

The exact model of Chinese SMEs credit assessment measures changed from traditional to mathematical, then turned to machine learning model, accompanied with information technology.

Machine learning avoids the limitation of non normality of sample data distribution in the traditional credit evaluation model, which can solve the noise and incomplete data, and shows excellent parallel processing ability. Machine learning model is characterized by large-scale parallel processing, strong robustness, great fault-tolerance, and powerful learning capacity. It has great application prospects and is the focus of research in the future, especially in aspects of algorithm design and parameter setting.

Up to now, many of the existing studies test several types of model on the evaluation of the loan credit risk of SMEs, but few compare them. One of the main goal of this study is to compare the effect of different machine learning models.

Due to the small sample size, this study mainly adopts some classical machine learning methods adapted to the dataset: Logistic Regression, Decision Tree, Random Forest, Naive Bayes, KNN, SVM and BP Neural Network, ADABOOST, XGBOOST.

We empirically test various credit evaluation models by using SMEs detailed loan data during 2011 to 2016, and find that the ensemble learning methods are often better than the other methods. XGBOOST model achieves the best result with a recall rate of 96.1% and a precision rate of 91.3%.

Due to the problem of unbalanced class and hidden incorrect label, we test random oversampling, random undersampling and three types of SMOTE sampling methods. We find that the SMOTE sampling methods are often better than random oversampling and random undersampling methods. In order to reduce the impact of label errors, we test the tri-training method and find that cooperated training can get a better result.

Lastly, we test the feature importance of each method mentioned above.

Although these methods are different, the most important characteristics of these methods are similar, which defines the most important influencing factors of default prediction, including remaining unpaid principal, floating interest rate, maturity time, actual interest rate and loan amount. Existing studies generally ignore the impact of floating interest rate and real interest rate on enterprise default probability.

5.2 Contributions and limitations

The innovation of this study mainly includes the following two aspects:

Firstly, we conduct a comprehensive test of various machine learning models. Compared with other studies that only use part of the models, this study is more comprehensive, and it is found that the ensemble learning method has a stable result on the whole.

Second, this study mainly uses the data of small and micro enterprises. Existing studies mainly use the data of small listed companies whichare not representative. This study solves the main problems existing in the dataset of small and micro enterprises, including category imbalance, label reliability and feature redundancy. Comparing to most previous studies only focusing on one or two of them, this study tests the potential impact of all these problems on the model results in a unified framework.

However, our research also has some deficiencies.

The first shortcoming of this study is that the sample used is only from one sub branch,so the regional representation is insufficient. Therefore, it is necessary to carefully evaluate the regional applicability of the results of this study. If the results of this study are applied to other banks, they should be adjusted. Secondly, if the credit evaluation model of small and micro enterprises is extended to other larger or smaller enterprises, more empirical research is needed. Third, due to the limited number of samples, this study does not use big data methods such as convolutional neural network and graph reasoning. In the future, the number of samples will be further expanded, and the big data method will be used for research to improve the universality of the model.

Different from developed countries, in China's market economy system, government policies have a great impact on the financing of small and micro enterprises by commercial banks. In the policy easing period, due to more supporting policies and financial subsidies, the credit status of small and micro enterprises will be significantly improved. The credit evaluation model of small and micro enterprises constructed in this study is mainly based on enterprise characteristics and financial performance, failingto include the policy environment into the scope of investigation.

In terms of research methods, this study uses machine learning method to mine the credit risk of small and micro enterprises hidden in data, and does not consider the internal logic and relevant risk characteristics of small and micro enterprise credit business of commercial banks, nor use the model to predict and compare with the actual situation. When using this model, commercial banks also need to make a comprehensive judgment in combination with the traditional credit evaluation model.

[This page is deliberately left blank.]

Bibliography

Abdelmoula, A. K. (2015). Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. *Accounting and Management Information Systems*, *14*(1), 79.

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, *35*(3), 1275–1292.

Abrahams, C. R., & Zhang, M. (2008).*Fair lending compliance: Intelligence and implications for credit risk management*(Vol. 13). John Wiley & Sons Press.

Aithal, V., & Jathanna, R. D. (2019). Credit risk assessment using machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 3482-3486.

Alam, N., Oliver, A., Denton, E. R. E., & Zwiggelaar, R. (2018). Automatic segmentation of microcalcification clusters. In M. Nixon, S. Mahmoodi, & R. Zwiggelaar (Eds), *Medical image understanding and analysis* (pp. 251 – 261). Springer International Publishing.

Alpaydın, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.*The Journal of Finance*,*23*(4), 589-609.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). Zeta analysis:A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance*, *1*(1), 29–54.

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, *48*(4), 733–755.

Bhattacharya, A., Biswas, S. K., & Mandal, A. (2022). Credit risk evaluation: a comprehensive study. *Multimedia Tools and Applications*, 1-51.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*,*5*, 1089–1105

Beringer, J., & Hüllermeier, E. (2006). Online clustering of parallel data streams. *Data & Knowledge Engineering*, *58*(2), 180–204.

Bhatia, K., Jain, H., Kar, P., Jain, P., & Varma, M. (2015). *Locally non-linear embeddings for extreme multi-label learning*. arXiv.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654.

Blum, A., & Mitchell, T. (1998, July 24-26). *Combining labeled and unlabeled data with co-training.*The 11th Annual Conference on Computational Learning Theory, New York, NY, United States.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Cai, X. T., & Yuan, Q.M. (2005).中小企业信用评级的指标体系与方法[Index system and method of credit rating of small and medium-sized enterprises].*Economic Management*,(11),46–49.

Chang, Z. (2015).*Construction of small and medium-sized enterprise credit evaluation system of bank a based on BP neural network*[Master's thesis]. Hunan University.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, J.H., Chen, H.M., & Ho, S.Y. (2005). Design of nearest neighbor classifiers: Multi-objective approach. *International Journal of Approximate Reasoning*, *40*(1), 3-22.

Chen, W. (2012).*Research on credit rating of small and medium-sized enterprises based on support vector machine method*[Master's thesis].Anhui University of Finance and Economics.

Chen, T., & Guestrin, C. (2016, August 13). *XGBoost: A scalable tree boosting system*.The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Franscisco, CA, United States.

Chi, G.T., & Li,H.X. (2019).基于逐步判别分析的小企业债信评级模型及实证[Debt rating model of small businesses and empirical analysis based on stepwise discriminant]. *Journal of Industrial Engineering and Engineering Management*,*33*(04),205–215.

Chuang, C.L., & Lin, R.H. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, *36*(2), 1685–1694.

Cooper, A. C., Gimeno-Gascon, F. J., & Woo, C. Y. (1991, August). *A resource-based prediction of new venture survival and growth*. Academy of management proceedings, Briarcliff Manor, NY, United States.

Cour, T., Sapp, B., & Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, *12*(42), 1501–1536.

Crosbie, P., & Bohn, J. (2018). Modeling default risk. InM. Crouhy, D. Galai & Z. Wiener (Eds)*World scientific reference on contingent claims analysis in corporate finance* (pp 471–506). World Scientific Publishing Company.

Dang, C., Li, Z., & Yang, C. (2018). Measuring firm size in empirical corporate finance. *Journal of Banking & Finance*, *86*, 159–176

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Duan, X.D. (2009).基于 BP 神经网络的中小企业信用评价模型研究[The model of mredit evaluation in small and medium sized enterprises based on BP neural network]. *Accounting and Finance*,(3),86–89.

Fan, M., Wang, Y.C., Zhang, Z.H., &Qiu, X.W. (2010).基于 AHP 法的中小企业信用评级模型研究[Research on credit rating model of small and medium-sized enterprises based on AHP]. *Communication of Finance and Accounting*,*06*,139–142.

Fan, B.L., &Zhu, W.B. (2003).中小企业信用评价指标的理论遴选与实证分析[An empirical study and choose of credit evaluation indexes of small-medium enterpries]. *Science Research Management*,(6),83–88.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), 368–378.

Freund, Y. (1999, June 27). *The alternating decision tree learning algorithm*. The 16th International Conference on Machine Learning, San Francisco, CA, United States.

Gao, J.G., Liu, X., & Zhu, C.C. (2015). 小微企业信用评估的数据挖掘方法综述[Summary of data mining methods for credit evaluation of small and micro enterprises]. *Financial Theory & Practice*, *10*, 98–101.

Gao, L.J. (2012).基于贝叶斯模型平均生存模型的中小企业信用风险估计[The estimation of credit risk of SMEs based on Bayesian model averaging survival model]. *Chinese Journal of Management Science*,*20*(S1),327–331.

Gao, M.X., Sun, Q.H., Hu, Q.& Zheng, F. (2022). 大中小微企业规模划型统计标准的实证研究[An empirical study on the statistical standard for the size classification of large,medium,small and micro-sized enterprises].*The Journal of Quantitative & Technical Economics*, 39(02), 164–181.

Gao, G., Wang, H., & Gao, P. (2021). Establishing a Credit Risk Evaluation System for SMEs Using the Soft Voting Fusion Model. *Risks*, *9*(11), 202.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review* *of Financial Studies*, *33*(5), 2223–2273.

Guo, N. (2013).政府?市场?谁更有效——中小企业融资难解决机制有效性研究[Government or market? Who is more effective].*Journal of Financial Research*,*03*,194–206.

Hady, M. F. A., & Schwenker, F. (2008, Dec 15-19). *Co-training by committee: A new semi-supervised learning framework*. 2008 IEEE International Conference on Data Mining Workshops, Pisa, Italy.

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble learning or deep learning? Application to default risk analysis. *Journal of Risk and Financial Management*, *11*(1), 12.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523–541.

Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, *14*(3), 515–516.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer

Hsieh, N.C., & Hung, L.P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, *37*(1), 534–545.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19.

Huang, Z., Chen, H., Hsu, C.J., Chen, W.H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, *37*(4), 543–558.

Huo, H.T. (2012).高科技中小企业信用风险指标体系及其评价方法[Research on evaluation index system of high-tech SME credit risk]. *Journal of Beijing Institute of Technology*（*Social Sciences Edition*）,*14*(01),60–65.

Hussein, A. S., Li, T., Yohannese, C. W., & Bashir, K. (2019). A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE.*International Journal of Computational Intelligence Systems*,*12*(2), 1412-1422.

Hwang, W. J., & Wen, K. W. (1998). Fast kNN classification algorithm based on partial distance search.*Electronics letters*, *34*(21), 2062-2063.

Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. *Cognitive Systems Research*, *52*, 317-324.

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research*, *5*(1), 1842-1845.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449.

Jin, W., Li, Z. J., Wei, L. S., & Zhen, H. (2000, Aug 21-25). *The improvements of BP neural network learning algorithm*. The 5th International Conference on Signal Processing, Beijing, China.

Jing,G., &Wang, H.L. (2013).基于模糊规则的中小企业信用评级系统研究[Small and medium enterprise credit rating system research based on fuzzy rule Algorithm]. *Heilongjiang Social Sciences*,(4),54–59.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010, December 13). *Discrimination aware decision tree learning*. IEEE International Conference on Data Mining, Sydney, Australia.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, *37*(9), 6233–6239.

Kirschenmann, K.(2016). Credit rationing in small firm-bank relationships. *Journal of financial Intermediation*, *26*, 68-99.

Klieštik, T., Kočišová, K., & Mišanková, M. (2015). Logit and probit model used for prediction of financial health of company. *Procedia Economics and Finance*, *23*, 850–855.

Krichene, A. (2017). Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank. *Journal of Economics, Finance and Administrative Science*, *22*(42), 3-24.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, *40*(13), 5125–5131.

Lamoureux, J. F., & Evans, T. A. (2011). Supply chain finance: a new means to support the competitiveness and resilience of global value chains.*Available at SSRN 2179944*.

Lando, D. (2009). Credit risk modeling. In *Handbook of Financial Time Series* (pp. 787-798). Springer, Berlin, Heidelberg.

Laurikkala, J. (2001, July 28). *Improving identification of difficult small classes by balancing class distribution.*Conference on Artificial Intelligence in Medicine in Europe, Berlin, Heidelberg, Germany.

Lee, J., & Yoon, T. (2017, Feb 19-22). *Analysis of relation between aging and telomere using datamining—Apriori, decision tree, and Support Vector Machine(SVM)*. The 19th International Conference on Advanced Communication Technology, Pyeongchang, Koera(South).

Lee, T.S., Chiu, C.C., Lu, C.J., & Chen, I.F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, *23*(3), 245–254.

Lee, T. S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *28*(4), 743–752.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113–1130.

Li, D., Yao, S., Liu, Y. H., Wang, S., & Sun, X. H. (2016, June). *Efficient design space exploration via statistical sampling and AdaBoost learning*. The 53nd ACM/EDAC/IEEE Design Automation Conference, IEEE, Austin, TX, USA.

Li, M., & Zhou, Z.H. (2007). Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *37*(6), 1088–1098.

Li,S.W., & Tong, G. R. (2015).Isomap-ABC-RVM 预测模型的构建及应用[Construction and application of Isomap-ABC-RVM prediction model].*Statistics & Decision*,(16),28–32.

Li, Z. J. (2017).微型企业信用评价指标体系的构建[Establishment of evaluation index system of credit state of mirco enterpries].*Technology Economics*,*36*(02),109–116.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest.*R news*,*2*(3), 18-22

Liu, P, & Shen, J. (2012).模糊综合评价法在中小企业信用评级中的应用[The application of fuzzy comprehensive evaluation method in the credit evaluation of small and medium enterprise]. *Science-Technology and Management*,*14*(06),51-54,59.

Liu, Y., Wang, Y., Ren, X., Zhou, H., & Diao, X. (2019). A classification method based on feature selection for imbalanced data. *IEEE Access*, *7*, 81794–81807.

Lu, H., Xu, Y., Ye, M., Yan, K., Gao, Z., & Jin, Q. (2019). Learning misclassification costs for imbalanced classification on gene expression data. *BMC Bioinformatics*, *20*(25), 681.

Lv, J.S. (2015).关于中小企业融资难、融资贵问题的思考[On financing constraints of small and medium enterprises]. *Journal of Financial Research*,(11),115–123.

Marqués, A. I., García, V., & Sánchez, J. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, *39*(11), 10244–10250.

McCulloch, R., & Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, *64*(1), 207–240.

Meng, B., Chi, G.T., & Gong, L. L. (2014).商户小额贷款信用评价模型[Credit evaluation model of mirco-credit of merchants]. *Science-Technology and Management*,*33*(12),103–108.

Meng, N.N., &Li, P. (2018). 中小微企业“麦克米伦缺口”成因及智能金融解决路径[Causes of Macmillan gap of small, medium-sized and micro enterprises and solutions of Intelligent Finance]. *South China Finance*, (7), 73–80.

Merton, R. C. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *The Journal of Finance*, *29*(2), 449–470.

Nanni, L., & Lumini, A. (2006). MppS: An ensemble of support vector machine based on multiple physicochemical properties of amino acids. *Neurocomputing*, *69*(13), 1688–1690.

Naseriparsa, M., Al-Shammari, A., Sheng, M., Zhang, Y., & Zhou, R. (2020). RSMOTE: Improving classification performance over imbalanced medical datasets. *Health Information Science and Systems*, *8*(1), 22.

National Bureau of Statistics.(2003) 统计上大中小型企业划分办法(暂行)[Categorization for Large, Medium and Small Enterprises in Statistics (Provisional)].*Beijing Statistics*, (6), 5.

Nguyen, N., & Caruana, R. (2008,August 24-27). *Classification with partial labels*. The 14th ACM SIGKDD International Conference on Knowledge discovery and Data mining, New York, NY, United States

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D., & Joksimovic, I. (2013). The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. *Expert Systems with Applications: An International Journal*, *40*(15), 5932–5944.

Niu,C. L. (2005).中小企业信用评级体系的构建[Construction of credit rating system for small and medium-sized enterprises].*Communication of Finance and Accounting*,(8),110–113.

Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science*, *42*(2), 150–165.

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications*, *41*(4, Part 2), 2052–2064.

Peng, W. (2012).我国上市中小企业信贷风险研究[Research on credit risk of Listed Small and medium-sized enterprises in China]. *Financial Regulation Research*,*01*,101–116.

Polikar, R. (2006). Ensemble based systems in decision making.*IEEE Circuits and systems magazine*,*6*(3), 21-45.

Qing, G., & Xin, C.Q. (2015).基于主成分-Logistic模型的中小企业信用评级研究——以大连市中小上市公司为例[Research on credit rating of small and medium-sized enterprises based on principal component logistic model]. *Journal of Jilin business and technology collage*,*30*(4),47-50,125.

Qiu, J., & Chen, J.S. (2014).中小企业信用评级指标体系构建的对策分析[Analysis on the construction of credit rating system for small and medium-sized enterprises]. *Commercial Economy*,(8),95–96.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Quinlan, J. R. (2014). *C4.5: Programs for machine learning*. Elsevier.

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007, June 20-24). *Self-taught learning: transfer learning from unlabeled data*,The 24th International Conference on Machine Learning, Corvalis, OR, United States.

Saaty, T. L. (2004). Decision making—the analytic hierarchy and network processes (AHP/ANP). *Journal of Systems Science and Systems Engineering*, *13*(1), 1–35.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674.

Satchidananda, S. S., & Simha, J. B. (2006). Comparing decision trees with logistic regression for credit risk analysis. *International Institute of Information Technology, Bangalore, India*.

Settles, B. (2012). Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114.

Shao, Y.H., Chen, W.J., Zhang, J.-J., Wang, Z., & Deng, N.-Y. (2014). An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recognition*, *47*, 3158–3167

Shen, Q. (2011).R*esearch on credit evaluation of small and medium-sized enterprises based on life cycle*[Master's thesis].Jiangsu University.

Singh, A., Wiktorsson, M., & Hauge, J. B. (2021). Trends In Machine Learning To Solve Problems In Logistics. *Procedia CIRP*, *103*, 67-72.

Songwattanasiri, P., & Sinapiromsaran, K. (2010, Dec 6-7). *Smoute: Synthetics minority over-sampling and under-sampling techniques for class imbalanced problem*. The Annual International Conference on Computer Science Education: Innovation and Technology, Special Track: Knowledge Discovery, Singapore.

Steijvers, T., Voordeckers, W., & Vanhoof, K. (2005, November 16). *The determinants of collateral: A decision tree analysis of SME loans*, RENT XIX Conference,Napoli, Italy .

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.

Sun, B., Chen, H., Wang, J., & Xie, H. (2018). Evolutionary under-sampling based bagging ensemble method for imbalanced data classification. *Frontiers of Computer Science*, *12*(2), 331–350.

Sun,W., & Wang, J.N. (2012).基于 AHP 的中小企业信用评级指标体系构建[Construction of credit rating index system of small and medium-sized enterprises based on AHP]. *Communication of Finance and Accounting*,*07*,19-21

Sun, Y., Tang, K., Zhu, Z., & Yao, X. (2018). Concept drift adaptation by exploiting historical knowledge.*IEEE transactions on neural networks and learning systems*,*29*(10), 4822-4832.

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293–300.

Tan, Q. M., Wu, J.K., & Zhao, L.M. (2009).基于 BP 神经网络的中小企业信用评价研究[Credit evaluation on SMEs Based on BP neural network]. *Journal of Northwest A&F University*(*Social Science Edition*),*9*(05),57-62.

Tsai, M.C., Lin, S.P., Cheng, C.C., & Lin, Y.P. (2009). The consumer loan default predicting model – An application of DEA–DA and neural network. *Expert Systems with Applications*, *36*(9), 11682–11690.

Uddin, M. S., Chi, G., Al Janabi, M. A., & Habib, T. (2022). Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance & Economics*, *27*(3), 3713-3729.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative review. *Journal of Machine Learning Research*, *10*(1), 66-71.

Wang, C., Li, H., Hao, Z., Li, X., Zou, C., Cai, P., Wang, Y., You, Y.Z., & Zhai, H. (2020).

Machine learning identification of impurities in the STM images. *Chinese Physics B*, *29*(11), 116805 1-5

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223-230.

Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic Regression Ensemble. *PLOS ONE*, *10*(2), e0117844.

Wang, J., Hedar, A.R., Wang, S., & Ma, J. (2012). Rough set and scatter search metaheuristic based feature selection for credit scoring. *Expert Systems with Applications*, *39*(6), 6123–6128.

Wang,L.M., Wang, J.R., & Li, X. (2016).基于集成学习的中小企业信用评价模型[Credit integration based learning evaluation model of small and medium-sized enterprises].*Modern Bankers*,*10*,136–137.

Wang, Q., &Yao, K. (2018). 机器学习方法在中小企业信用评估中的应用研究[Application of machine learning method in credit evaluation of small and medium-sized enterprises].*Special Zone Economy*, *01*, 145–147.

Wang, X., Wang, Y., &Chen, J.D. (2021). 我国中小微企业信用评价研究现状与发展趋势[Research status and development trend of credit evaluation of small, medium and micro enterprises in China]. *Credit Reference*, *39*(05), 62–70.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data. *Procedia Computer Science*, *174*, 141-149.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, *32*(10), 2543–2559.

Wu,Y. (2013).基于神经网络的中小企业信用评价体系的构建[Construction of credit evaluation system of small and medium-sized enterprises based on neural network]. *The Science Education Article Collects*,*06*,89-91,93

Xia, H. (2019).基于支持向量机回归集成的小微企业信用风险度评估模型研究[Study on the micro and small enterprises cedit risks evaluation model based on support-vector machine regression ensemble]. *Credit*,04,21-27

Xia, Y.C. (2013).*Research on credit evaluation model and application of small and medium-sized enterprises based on support vector machine*[Master’s thesis]. Central South University.

Xiong, B.Y., Wang, G.Y., & D, W.B. (2016).基于样本权重的不平衡数据欠抽样方法[Under-sampling method basedonsample weight for imbalanced data]. *Journal of Computer Research and Development*,*53*(11),2613–2622.

Xu, X.P., & Ma, W.J. (2011).非上市中小企业贷款违约率的定量分析——基于判别分析法和决策树模型的分析[Quantitative analysis of loan default rate of unlisted small and medium-sized enterprises]. *Journal of Financial Research*,*03*,111–120.

Yang, Z.J., Zhang, L.M., & Xu, X.J. (2011).小企业贷款信用评级体系创新研究[Research on the innovation of small enterprise loan credit rating system]. *Zhejiang Finance*,*10*,71–73.

Ye, X.F., & Lu, Y.H. (2017).基于随机森林融合朴素贝叶斯的信用评估模型[Credit evaluation model based on random forest fusion naive Bayes]. *Mathematics in Practice and Theory*,*47*(2),68–73.

Yeh, C.C., Lin, F., & Hsu, C.Y. (2012). A hybrid KMV model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, *33*, 166–172.

Yi, S.Q. (2007).*Research on credit behavior of cluster financing of small and medium-sized enterprises*[Master’s thesis]. Central South University

Yu, L., Wang, S., & Lai, K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, *34*(2), 1434–1444.

Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*, *37*(2), 1351-1360.

Zhang, C., Guo, J., & Lu, J. (2019). Research on classification method of high-dimensional class-imbalanced datasets based on SVM.*International Journal of Machine Learning and Cybernetics*,*10*(7), 1765-1778.

Zhang, M. L., & Yu, F. (2015, June25–31). *Solving the partial label learning problem: An instance-based approach*. The24thInternational Joint Conference on Artificial Intelligence,Palo Alto, CA, USA.

Zhang, M. L., Zhou, B. B., & Liu, X. Y. (2016, August 13). *Partial label learning via feature-aware disambiguation*.The 22nd ACM SIGKDD international conference on knowledge discovery and data mining, NY,USA.

Zhang, X.L. (2014). Heuristic ternary error-correcting output codes via weight optimization and layered clustering-based approach.*IEEE Transactions on Cybernetics*,*45*(2), 289-301.

Zhang, H. (2008).基于供应链金融的中小企业信用评级模型研究[Research on credit rating model of small and medium-sized enterprises based on supply chain finance]. *Journal of Southeast University(Philosophy and Social Science)*,*10*(S2),54–58.

Zhang,H.Y., & Li, H. (2017).BP 神经网络模型在中小企业融资征信评估中的应用[Application of BP neural network model in financing credit investigation and evaluation of small and medium-sized enterprises].*Market Modernization*,(18),172–173.

Zhang, L., Yang, Z.S., & Chen, S. (2004).KMV 模型在上市公司信用风险评价中的应用研究[An application of KMV model in credit risk evaluation of listed companies]. *Systems Engineering*,(11),84–89.

Zheng, X.J. (2015).互联网金融背景下科技型中小企业信用评级实证研究[An empirical study on small and medium-sized technological enterprise credit rating in the background of internet finance]. *Credit Reference*,*33*(09),54–58.

Zhong, T.L. & Jia, L.H. (2005).中小企业信用评价的神经网络法[Neural network method for credit evaluation of small and medium-sized enterprises]. *Journal of Technical Economics & Management*,*5*,30–32.

Zhou,Z. H., & Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*,*17*(11), 1529-1541.

Zhu, H., Zhang, H.M., & Xu, C. (2015).基于相对熵的存货质押融资模式下中小企业信用评价[Credit Evaluation of SEMs in Inventory Financing Mode based on Relative Entropy]. *Journal of Guizhou University of Engineering Science*,*33*(02),132–138.