



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Loan Review Mechanism of Online Credit Loan Products: The case of Tax E Financing of Jiangsu Bank

YUE Zhenhua

Doctor of Management

Supervisors

PhD Pedro Inacio, Assistant Professor,

ISCTE University Institute of Lisbon

PhD MA Yongkai, Professor

University of Electronic Science and Technology of China

October, 2022



**BUSINESS
SCHOOL**

Marketing, Operations and General Management Department

Loan Review Mechanism of Online Credit Loan Products: The case of Tax E Financing of Jiangsu Bank

YUE Zhenhua

Doctor of Management

Supervisors

PhD Pedro Inacio, Assistant Professor,

ISCTE University Institute of Lisbon

PhD MA Yongkai, Professor

University of Electronic Science and Technology of China

October, 2022



**BUSINESS
SCHOOL**

Marketing, Operations and General Management Department

Loan Review Mechanism of Online Credit Loan Products: The case of Tax E
Financing of Jiangsu Bank

YUE Zhenhua

Doctor of Management

Jury:

PhD José Manuel Gonçalves Dias, Associate Professor with Habilitation

ISCTE University Institute of Lisbon

PhD Luís Pedro Vilela Pimentel, Assistant Professor,

Universidade Europeia

PhD Li Ping, Professor,

Electronic Science and Technology of China

PhD Pedro Manuel de Sousa Leite Inácio, Retired Professor,

ISCTE University Institute of Lisbon

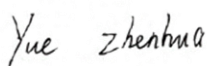
PhD Vasco Barroso Gonçalves, Assistant Professor,

ISCTE University Institute of Lisbon

October, 2022

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.

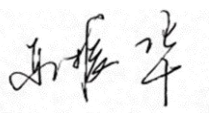
Signed: 

Date: 2022.10.21

Name: YUE Zhenhua

作者申明

本人郑重声明：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: 

日期: 2022.10.21

姓名(拼音): YUE Zhenhua

[This page is deliberately left blank.]

Abstract

This thesis takes Jiangsu Bank's "Tax E Financing" product as the analysis object, trying to improve its loan review mechanism. The basic features, financial features and tax features of defaulting enterprises, non-defaulting enterprises and non-passing enterprises are collected and calculated. Through statistical analysis and machine learning algorithms, a credit risk evaluation index system is constructed and the best credit risk assessment method is chosen. The main conclusions are as follows.

First, this thesis finds that there are obvious differences in characteristics between defaulting enterprises and non-defaulting enterprises, and some of the characteristics of non-passing enterprises are better than those of non-defaulting enterprises, indicating that they have a high probability of being rejected by mistake.

Second, it is found that "tax features" and "basic features + tax features" have the best classification performance; "financial features" have the worst classification performance. random forest algorithm is used to optimize the feature sets of "basic features" and "basic features + taxation features", and it is found that after eliminating the unimportant features, the classification ability of feature set can be effectively improved.

Third, based on the optimized "tax feature" and "basic feature + tax feature", this thesis uses various algorithms of supervised learning and semi-supervised learning to train and test the classifier. The study found that the classifier trained by the random forest algorithm has the best performance, with an accuracy rate of 98.89%. On the premise of not judging any non-defaulting company, it can identify 89.47% of the defaulting companies.

Keywords: Tax E Financing, credit risk, index system, machine learning, classification algorithm

JEL: G21, P34

[This page is deliberately left blank.]

Resumo

Esta tese tem como objeto de análise o produto “Tax E Financing” do Jiangsu Bank, tentando melhorar o seu mecanismo de revisão de empréstimos. Através de uma análise estatística e algoritmos de aprendizado de máquina, é construído um sistema de índice de avaliação de risco de crédito e escolhido o melhor método de avaliação de risco de crédito. As principais conclusões são as seguintes.

Em primeiro lugar, esta tese constata que existem diferenças óbvias nas características entre as empresas inadimplentes e as empresas não inadimplentes, e algumas das características das empresas inadimplentes são melhores do que as das empresas não inadimplentes, indicando que elas têm uma alta probabilidade de serem rejeitadas por engano.

Em segundo lugar, verifica-se que “características fiscais” e “características básicas + características fiscais” têm o melhor desempenho de classificação; “características financeiras” têm o pior desempenho de classificação. O algoritmo de floresta aleatória é usado para otimizar os conjuntos de recursos de “características básicas” e “recursos básicos + recursos de tributação”, e verifica-se que após eliminar os recursos sem importância, a capacidade de classificação do conjunto de recursos pode ser efetivamente melhorada.

Terceiro, com base no “recurso fiscal” e no “recurso básico + recurso fiscal” otimizados, este trabalho utiliza vários algoritmos de aprendizado supervisionado e aprendizado semi-supervisionado para treinar e testar o classificador. O estudo constatou que o classificador treinado pela floresta aleatória apresenta o melhor desempenho, com índice de acerto de 98,89%, com a premissa de não julgar nenhuma empresa inadimplente, consegue identificar 89,47% das empresas inadimplentes.

Palavras-chave: Tax E Financing, risco de crédito, sistema de indexação, aprendizagem de máquina, algoritmo de classificação

JEL: G21, P34

[This page is deliberately left blank.]

摘要

本文以江苏银行的“税 e 融”产品为分析对象，试图改进其审贷机制。为此，本文收集和计算了通过“税 e 融”申请贷款的违约企业、未违约企业和未通过企业的基础特征、财务特征和纳税特征，通过统计分析和机器学习算法，为“税 e 融”构建了信用风险评价指标体系，并进行信用风险评价方法的选择。主要结论如下。

第一，本文发现违约企业与未违约企业具有明显的特征差异，而未通过企业的部分特征优于未违约企业，表明其被错误拒绝的概率很高。

第二，本文发现“纳税特征”和“基础特征+纳税特征”具有最好的分类性能；“财务特征”的分类性能最差。同时，本文利用随机森林算法对“基础特征”和“基础特征+纳税特征”特征集进行了优化，发现剔除不重要性特征后，能有效提升数据的分类能力。

第三，本文基于优化后的“纳税特征”和“基础特征+纳税特征”，分别利用监督学习和半监督学习的各类算法进行分类器的训练和测试。研究发现，采用随机森林算法训练出的分类器具有最好的性能，准确率达到 98.89%，在不错判任何一个未违约企业的前提下，能识别出 89.47%的违约企业。

关键词：税 e 融，信用风险，指标体系，机器学习，分类算法

JEL: G21, P34

[This page is deliberately left blank.]

Acknowledgements

I would like to gratefully thank my supervisors Prof. Pedro Inacio and Prof. Ma Yongkai for their intellectual guidance and support, enormous patience and encouragement, and helpful comments and suggestions throughout the course of this thesis. I would like to thank the team members of Prof. Ma. They gave me great help in software usage. Additionally, I would also like to thank my family for their unconditional love and support.

[This page is deliberately left blank.]

致谢

衷心感谢我的导师 **Pedro Inacio** 教授和马永开教授在整个论文过程中给予的指导和
支持，感谢他们对我的耐心和鼓励，以及对论文提出的宝贵意见和建议。感谢马永开教
授的团队成员，他们在软件使用方面给了我很大的帮助。此外，我还要感谢我的家人，
谢谢他们对我无条件的爱和支持。

[This page is deliberately left blank.]

Contents

Chapter 1: Introduction	1
1.1 Research background	1
1.2 Current loan review mechanism of “Tax E Financing”	8
1.3 Research problems	10
1.4 Research questions	13
1.5 Research structure	14
1.6 Research value and expected contribution	15
Chapter 2: Literature Review	17
2.1 Concept and theoretical basis of credit risk	17
2.1.1 Credit and credit risk	17
2.1.2 Information asymmetry theory	19
2.2 Research on enterprise credit risk assessment methods	22
2.2.1 Qualitative credit risk assessment methods	22
2.2.2 Quantitative credit risk assessment methods	23
2.3 Research on credit risk assessment index system	35
2.3.1 The application of hard information in credit evaluation	36
2.3.2 The application of soft information in credit evaluation	39
2.4 Research on information sharing	55
2.4.1 Theoretical research on the impact of information sharing	55
2.4.2 Empirical research on the impact of information sharing	57
2.4.3 Research on China’s credit information sharing	58
2.5 Literature summary	59
Chapter 3: Research Methods	61
3.1 Model evaluation methods	61
3.1.1 Empirical error and generalization error	61
3.1.2 Performance measurement	62
3.2 Feature processing methods	65
3.2.1 Principal component analysis	65
3.2.2 Feature selection	66
3.3 Classification algorithms	66

3.3.1 Support Vector Machine	66
3.3.2 Random Forest	69
3.3.3 Logistic Regression	72
3.3.4 BP Neural Network	74
3.3.5 Transductive Support Vector Machine	75
Chapter 4: Analysis on the Credit Evaluation Index System of “Tax E Financing”	77
4.1 Sample data collection and processing	78
4.1.1 Sample data collection	78
4.1.2 Generating new features	79
4.2 Feature analysis of sample enterprises	81
4.2.1 Analysis on the features of defaulting enterprises	81
4.2.2 Analysis on the features of non-passing enterprises	87
4.3 Chapter summary	91
Chapter 5: Optimization of the Credit Evaluation Index System of “Tax E Financing”	93
5.1 Analysis of the predictive ability of tax features	93
5.1.1 Feature engineering	93
5.1.2 Data visualization	95
5.1.3 Predictive ability comparison of SVM classifiers on 7 sub-samples	99
5.1.4 Summary	107
5.2 Credit evaluation index system optimization based on reliable information	108
5.2.1 Optimization of tax features	109
5.2.2 Optimization of basic features and tax features	110
5.3 Chapter summary	112
Chapter 6: Credit evaluation Method Selection of “Tax E Financing”	113
6.1 SVM	114
6.1.1 Training result of SVM classifier on tax features	114
6.1.2 Training result of SVM classifier on tax features and basic features	115
6.1.3 Performance comparison before and after feature index optimization	115
6.2 LR	116
6.2.1 Training result of LR classifier on tax features	117
6.2.2 Training result of LR classifier on tax features and basic features	117
6.3 BPNN	118
6.3.1 Training result of BPNN classifier on tax features	118
6.3.2 Training result of BPNN classifier on tax features and basic features	118
6.4 RF	119

6.4.1 Training result of RF classifier on tax features	119
6.4.2 Training result of RF classifier on tax features and basic features.....	119
6.5 TSVM.....	120
6.5.1 Training result of TSVM classifier on tax features	120
6.5.2 Training result of TSVM classifier on tax features and basic features	121
6.6 Performance comparison of classifiers trained by different algorithms.....	121
6.7 Chapter summary	124
Chapter 7: Conclusions, Suggestions and Limitations.....	127
7.1 Research conclusions	127
7.2 Policy suggestions	128
7.3 Research limitations	129
Bibliography.....	131
Webliography	143
Annex A: Tables	145
Annex B: Figures	155

[This page is deliberately left blank.]

List of Tables

Table 1.1 Statistical criteria for the classification of large, medium, small and micro enterprises	2
Table 1.2 Evaluating indicator system of tax credit rating and their scores.....	6
Table 1.3 Loan review rule of “Tax E Financing”	9
Table 1.4 Default information of “Tax E Financing” and offline SMEs loan business of Nantong Branch in the second half of 2017.....	11
Table 1.5 Approval information of “Tax E Financing” and offline SMEs loan business of Nantong Branch in the second half of 2017	11
Table 2.1 Summary of the credit risk evaluation index system for SMEs	53
Table 4.1 Industry distribution of different samples	79
Table 4.2 Added financial features and calculation formulas	80
Table 4.3 Balance sheets comparison of defaulting and non-defaulting samples	83
Table 4.4 Income statement comparison of defaulting and non-defaulting samples	84
Table 4.5 Financial ratios comparison of defaulting and non-defaulting samples	85
Table 4.6 Tax features comparison of defaulting and non-defaulting samples	86
Table 4.7 Top 10 rules triggered and their percentages.....	88
Table 4.8 Comparison of the basic features of the three types of samples.....	89
Table 4.9 Comparison of the tax features of the three types of samples	90
Table 5.1 Feature importance of sample 3	110
Table 5.2 Feature importance of sample 5	111
Table 6.1 The optimized credit evaluation index system	113

[This page is deliberately left blank.]

List of Figures

Figure 3.1 ROC curve and AUC	64
Figure 5.1 2D and 3D visualization of sample 1	95
Figure 5.2 2D and 3D visualization of sample 2	96
Figure 5.3 2D and 3D visualization of sample 3	96
Figure 5.4 2D and 3D visualization of sample 4	97
Figure 5.5 2D and 3D visualization of sample 5	97
Figure 5.6 2D and 3D visualization of sample 6	98
Figure 5.7 2D and 3D visualization of sample 7	99
Figure 5.8 Performance comparison of SVM classifier on 7 sub-samples	108
Figure 6.1 Performance comparison for samples before and after feature optimization	116
Figure 6.2 Performance comparison of five algorithms on the new sample 3	122
Figure 6.3 Performance comparison of five algorithms on the new sample 5	124

[This page is deliberately left blank.]

List of Abbreviations

ANN	Artificial Neural Network
AHP	Analytic Hierarchy Process
AUC	Area under ROC Curve
BPNN	Back Propagation Neural Network
CPN	Credit Portfolio View
DES	Dynamic Ensemble Selection
DT	Decision Tree
EBIT	Earnings before Interest and Tax
FPR	False Positive Rate
FA-SV	Firefly Algorithm Support Vector Machine
FS-MRI	Feature Selection-Multiple Ranking Information
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity
GBDT	Gradient Boosting Decision Tree
GDP	Gross Domestic Product
LightGBM	Light Gradient Boosting Machine
k-NN	k-Nearest Neighbors
LR	Logit Regression,
MLP	Multi-layer Perceptron
PCA	Principal Component Analysis
P2P	Peer to Peer
RF	Random Forest
ROC	Receiver Operating Characteristic
RT	Regression Tree
SCSRF	Syncretic Cost-sensitive Random Forest
SMEs	Small and Medium-Sized Enterprises
SVM	Support Vector Machine
SVME-AIR	SVM Ensemble Model with Artificial Imbalance Rate
TPR	True Positive Rate
TSVM	Transductive Support Vector Machine
VAT	Value Added Tax

VSG	Virtual Sample Generation
XGBoost	eXtreme Gradient Boosting

Chapter 1: Introduction

1.1 Research background

In many countries, enterprises are classified into different scale levels according to indicators such as the number of employees, total assets, and/or operating income (i.e., Total Revenues), so as to formulate targeted policies. Different countries have different classification criteria. In China, classification criteria for enterprise scale levels have undergone several adjustments. The earliest criteria were promulgated in the 1950s. Then the number of employees was used to classify enterprises into three scale levels: large, medium and small. The enterprises with more than 3,000 employees were large, the enterprises with 500-3,000 employees medium-sized, and the ones with less than 500 employees small. With the adjustment of the economic structure, more and more micro-enterprises are emerging, which play an important role in China's economic development and therefore attract more attention of the central government.

In September 2011, the Chinese Ministry of Industry and Information Technology, the National Bureau of Statistics, the National Development and Reform Commission and the Ministry of Finance jointly issued the “Notice on Printing and Distributing Standards for SMEs (Small and Medium-sized Enterprises)”, and formulated the “Statistical Classification Criteria for Large, Medium, Small and Micro Enterprises”. The scale levels of enterprises were then changed from the previous “large, medium and small” to “large, medium, small and micro”. For the first time, micro-enterprises were classified as a single category. However, the National Bureau of Statistics abolished that document on December 28, 2017. The current implemented document is the “Notice on Printing and Distributing the ‘Statistics Classification Criteria for Large, Medium, Small and Micro Enterprises (2017)’”. The latest classification criteria are shown in Table 1.1.

With the support of a series of policies of the central government, China's micro and small enterprises have developed rapidly and played a more important role in the national economy. Especially in the context of the downward pressure on China's economy in recent years and the arduous task of addressing overcapacity, the rapid development of micro and small enterprises has become an important supporting force for maintaining and enhancing China's employment ratio. More importantly, the emergence and development of technological innovative micro and

Table 1.1 Statistical criteria for the classification of large, medium, small and micro enterprises

Industry	Indicators	Unit	Large	Medium	Small	Micro
Agriculture, forestry, animal husbandry and fishery	Operating income (Y)	10 thousands RMB	$Y \geq 20,000$	$500 \leq Y < 20,000$	$50 \leq Y < 500$	$Y < 50$
	Number of employees (X)	# of persons	$X \geq 1,000$	$300 \leq X < 1,000$	$20 \leq X < 300$	$X < 20$
Industry*	Operating income (Y)	10 thousands RMB	$Y \geq 40,000$	$2,000 \leq Y < 40,000$	$300 \leq Y < 2,000$	$Y < 300$
	Operating income (Y)	10 thousands RMB	$Y \geq 80,000$	$6,000 \leq Y < 80,000$	$300 \leq Y < 6,000$	$Y < 300$
Construction industry	Total assets (Z)	10 thousands RMB	$Z \geq 80,000$	$5,000 \leq Z < 80,000$	$300 \leq Z < 5,000$	$Z < 300$
	Number of employees (X)	# of persons	$X \geq 200$	$20 \leq X < 200$	$5 \leq X < 20$	$X < 5$
Wholesale industry	Operating income (Y)	10 thousands RMB	$Y \geq 40,000$	$5,000 \leq Y < 40,000$	$1,000 \leq Y < 5,000$	$Y < 1,000$
	Number of employees (X)	# of persons	$X \geq 300$	$50 \leq X < 300$	$10 \leq X < 50$	$X < 10$
Retail industry	Operating income (Y)	10 thousands RMB	$Y \geq 20,000$	$500 \leq Y < 20,000$	$100 \leq Y < 500$	$Y < 100$
	Number of employees (X)	# of persons	$X \geq 1,000$	$300 \leq X < 1,000$	$20 \leq X < 300$	$X < 20$
Transportation industry *	Operating income (Y)	10 thousands RMB	$Y \geq 30,000$	$3,000 \leq Y < 30,000$	$200 \leq Y < 3,000$	$Y < 200$
	Number of employees (X)	# of persons	$X \geq 200$	$100 \leq X < 200$	$20 \leq X < 100$	$X < 20$
Warehousing industry *	Operating income (Y)	10 thousands RMB	$Y \geq 30,000$	$1,000 \leq Y < 30,000$	$100 \leq Y < 1,000$	$Y < 100$
	Number of employees (X)	# of persons	$X \geq 1,000$	$300 \leq X < 1,000$	$20 \leq X < 300$	$X < 20$
Postal industry	Operating income (Y)	10 thousands RMB	$Y \geq 30,000$	$2,000 \leq Y < 30,000$	$100 \leq Y < 2,000$	$Y < 100$
	Number of employees (X)	# of persons	$X \geq 300$	$100 \leq X < 300$	$10 \leq X < 100$	$X < 10$
Accommodation industry	Operating income (Y)	10 thousands RMB	$Y \geq 10,000$	$2,000 \leq Y < 10,000$	$100 \leq Y < 2,000$	$Y < 100$
	Number of employees (X)	# of persons	$X \geq 300$	$100 \leq X < 300$	$10 \leq X < 100$	$X < 10$
Catering industry	Operating income (Y)	10 thousands RMB	$Y \geq 10,000$	$2,000 \leq Y < 10,000$	$100 \leq Y < 2,000$	$Y < 100$
	Number of employees (X)	# of persons	$X \geq 2,000$	$100 \leq X < 2,000$	$10 \leq X < 100$	$X < 10$

Information transmission industry *	Operating income (Y)	10 thousands RMB	$Y \geq 100,000$	$1,000 \leq Y < 100,000$	$100 \leq Y < 1,000$	$Y < 100$
Software and information technology services industry	Number of employees (X)	# of persons	$X \geq 300$	$100 \leq X < 300$	$10 \leq X < 100$	$X < 10$
	Operating income (Y)	10 thousands RMB	$Y \geq 10,000$	$1,000 \leq Y < 10,000$	$50 \leq Y < 1,000$	$Y < 50$
Real estate development and operation industry	Operating income (Y)	10 thousands RMB	$Y \geq 200,000$	$1,000 \leq Y < 200,000$	$100 \leq Y < 1,000$	$Y < 100$
	Total assets (Z)	10 thousands RMB	$Z \geq 10,000$	$5,000 \leq Z < 10,000$	$2,000 \leq Z < 5,000$	$Z < 2,000$
Property management industry	Number of employees (X)	# of persons	$X \geq 1,000$	$300 \leq X < 1,000$	$100 \leq X < 300$	$X < 100$
	Operating income (Y)	10 thousands RMB	$Y \geq 5,000$	$1,000 \leq Y < 5,000$	$500 \leq Y < 1,000$	$Y < 500$
Leasing and business services industry	Number of employees (X)	# of persons	$X \geq 300$	$100 \leq X < 300$	$10 \leq X < 100$	$X < 10$
	Total asset (Z)	10 thousands RMB	$Z \geq 120,000$	$8,000 \leq Z < 120,000$	$100 \leq Z < 8,000$	$Z < 100$
Other unspecified industries *	Number of employees (X)	# of persons	$X \geq 300$	$100 \leq X < 300$	$10 \leq X < 100$	$X < 10$

Notes: 1. Large, medium and small enterprises must meet the lowest limit of the listed indicators at the same time, otherwise they will be classified into the smaller level, while micro enterprises only need to meet one of the listed indicators. 2. The scope of each industry in the attached table is subject to the National Economic Industry Classification (GB/T4754-2017). The items with * are industry portfolio categories. Industry includes mining, manufacturing, electricity, heat, gas and water production and supply. Transportation industry includes road transport, water transport, air transport, pipeline transport, multimodal transport and transportation agency, loading and unloading, excluding railway transportation. Warehousing industry includes general warehousing, low temperature storage, dangerous goods warehousing, grain, cotton and other agricultural products warehousing, Chinese medicinal materials warehousing and other warehousing. Information transmission industry includes telecommunications, radio and television and satellite transmission services, internet and related services.

Source: "Notice on Printing and Distributing the 'Statistics Classification Criteria for Large, Medium, Small and Micro Enterprises (2017)'" issued by the National Bureau of Statistics.

small enterprises have become a great boost to China's stable economic growth and economic restructuring. Statistics from the State Administration for Industry and Commerce show that by the end of 2017, there were about 28 million micro, small and medium-sized enterprises in the country, and more than 65 million individual business, accounting for more than 90% of the total number of market entities. "The China Micro and Small Enterprise Financial Services Report (2018)" jointly issued by the People's Bank of China and China Bank and Insurance Supervision and Management Commission shows that China's micro, small and medium-sized enterprises contribute to more than 50% of tax revenue, more than 60% of GDP, more than 70% of technology innovation and more than 80% of urban employment, which are important carriers for mass entrepreneurship and innovation.

However, the strong financing constraints and huge financing gaps that Chinese micro, small and medium-sized enterprises faced are mismatched with their proportions and contributions to the Chinese economy. The World Bank Enterprise Surveys (WBES) proposes that financing difficulties are the primary factor constraining the development of SMEs, and this problem is even more pronounced in China. The data shows that the average life expectancy of SMEs in China is about 3 years, while the average lives of SMEs in the United States and Japan are 8 years and 12 years, respectively. Financing constraints of micro, small and medium-sized enterprises are reflected in two aspects: narrow financing channels and high financing costs.

Firstly, in terms of financing channels, micro, small and medium-sized enterprises rely mainly on internal financing (i.e. internally generated funds); their external financing channels are narrow, especially while the credit resources are crowded by large enterprises. According to the statistics of China Small and Medium Enterprises Association, at the end of 2015, among the more than 20 million SMEs surveyed, only about 6 million enterprises obtained bank loans, and nearly 75% of SMEs failed to obtain formal bank credit. As of the end of 2018, the balance of corporate loans for micro and small enterprises in China was 26 trillion RMB, accounting for only 32.1% of all corporate loans. At the same time, only a few companies that had obtained bank loans could get the expected financing quotas, and most of the financing needs could not be met. In the absence of insufficient internal financing sources and insufficient bank credit supports, many micro, small and medium-sized enterprises rely on microfinance companies and internet loans to meet their financing needs.

Secondly, in terms of financing costs, the comprehensive financing costs of micro, small and medium-sized enterprises are much higher than those of large enterprises. The comprehensive financing cost not only includes the cost of capital use, which is the financing

interest, but also includes the cost of capital acquisition, such as guarantees fee, evaluation fee and registration fee. In February 2018, the China Finance Research Center of the School of Economics and Management of Tsinghua University announced the China Social Financing Cost Index. The data shows that the average cost of social financing for enterprises in China at that time was 7.6%. The average cost of microfinance companies and internet loans was 21.9% and 21%, respectively, much higher than the average cost of bank loans of 6.6%. The above costs only reflect the cost of capital use, and do not include the cost of capital acquisition. For micro and small enterprises, they also need to pay about 1%-2% guarantee fee, and 0.1%-1% evaluation, registration and other handling fees.

The reasons for the financing difficulty not only lie in the internal factors of micro and small enterprises, but also in the outside factors such as the market and system, which are mainly manifested in the following two aspects.

First, from the perspective of operating costs of banks, the four characteristics that short-term, frequent, small, and urgent credit needs make the micro and small enterprises excluded by the scale economies of the financial industry. Its “short-term and frequent” is that the credit loans of micro and small enterprises are mainly used for operating expenses, such as purchasing raw materials, not for investment expenditures. Therefore, it can be quickly refunded, but needs to be turned (rolled) over from time to time. Its “small” lies in the small amount of credit demand caused by the small-scale production and operation. Meanwhile, as their production and operation cycle are difficult to predict, and the capital demands are difficult to plan, the financing demands of micro and small enterprises are often sudden and urgent. According to empirical data, the average financing amount of each SMEs is 0.5% of that of large enterprises, but their average financing frequency is 5-6 times the one of large enterprises. Therefore, it will greatly increase the operating and supervision costs of banks, thus reducing their willingness to lend. Therefore, credit resources pursue large-scale and low-cost, and are more willing to allocate funds to large enterprises with large capital demand but relatively low frequency.

Second, from the perspective of credit risk, the high business risk and moral hazard of micro and small enterprises have also aggravated the bank’s adverse choice. Micro and small enterprises have higher operational risks. Due to their small scale, insufficient funds, low technical level, unsound management system and unstable profitability, micro and small enterprises lack the ability to resist macroeconomic risks, and are more likely to be eliminated by the market. What is more important is that the risk level of micro and small enterprises is difficult to identify by banks. On the one hand, it is difficult for micro and small enterprises to provide standardized business and financial information. Due to their imperfect internal

management system, some enterprises do not even have financial statements. Even if financial statements are available, the information may be seriously distorted. On the other hand, micro and small enterprises have difficulty in providing banks with effective loan collateral due to the lack of assets, especially the lack of fixed assets. Therefore, it is necessary for banks to spend a large amount of collection and investigation costs to accurately evaluate the financial status of micro and small enterprises, which is one of the main reasons for reducing their willingness of credit supply. The information asymmetry between micro and small enterprises and banks may lead to the after-the-fact moral hazard of enterprises. For example, the credit funds may be used for high-risk projects, and even used to meet the consumption needs of business owners, which will increase the credit risk of banks. Meanwhile, it will also exacerbate the bank's reverse choices. Because banks cannot accurately identify the credit risks of every micro and small enterprise, they may choose to substantially increase the average financing costs, or reduce the financing supply to all enterprises. This will cause micro and small-enterprises with low risk to abandon credit financing and seek other low-cost financing channels, so that all the credit funds are obtained by high-risk enterprises, leading to the phenomenon of "bad money expelling good money".

In summary, one of the main reasons for the financing difficulty of micro and small enterprises is that it is hard for banks to identify the credit risks of those enterprises, which is mainly caused by the lack of reliable information that can accurately reflect the operation and financial status of micro and small enterprises. In order to solve the financing problems of micro and small enterprises, in 2015, the State Administration of Taxation and the China Banking Regulatory Commission jointly launched the work of "Interaction between Banks and Tax Authorities" to help the development of micro and small enterprises. The essence of this work lies in the fact that local tax authorities share the historical tax information of micro and small enterprises with banks, and convert their "tax credits" into "bank credits", providing a decision-making basis for banks to allocate loans. It can not only improve the income of the bank, but also meet the financing demands of micro and small enterprises and increase their enthusiasm to pay tax, which will achieve a win-win situation for enterprises, banks and government.

Table 1.2 Evaluating indicator system of tax credit rating and their scores

Evaluating indicators	Score
(1)Tax registration	
1. Registration of starting operation	
2. Registration of tax withholding	
3. Registration of tax changes	15'
4. Use of registration document	
5. Annual inspection and renewal	

6. Report of bank account	
7. Status of tax recognition	
(2)Tax declaration	
1. Rate of scheduled tax declarations	
2. Accuracy rate of scheduled tax declarations	
3. Rate of scheduled withholding and paying tax	25'
4. Accuracy rate of scheduled withholding and paying tax	
5. Submitting financial accounting statements and other tax information	
(3)Management of account book and certificate	
1. Submitting financial accounting system or financial accounting treatment methods and accounting software	
2. Setting up, keeping accounts, and vouchers according to regulations, and performing accounting according to legal and valid vouchers	15'
3. Custody, issuance, use, and acquisition of invoices	
4. Installation, storage and use of tax control devices and anti-counterfeiting tax control systems.	
(4)Tax payment	
1. Rate of tax paid on time	25'
2. No Unpaid tax	
3. Rate of withholding and paying tax paid on time	
(5)Violation of tax laws and administrative regulations	
1. No Tax-related crime record	20'
2. No Tax administrative penalty record	
3. No Other tax violation record	

Source: "Tax Credit Management Measures (for Trial Implementation)" issued by the State Administration of Taxation.

Tax credit refers to a kind of credit that taxpayers (the enterprises) fulfill their tax obligations according to law and is generally recognized by the society. The tax credit rating is rated by the tax authorities on the basis of the implementation of tax obligations of taxpayers, according to the "Trial measures for Tax Credit Rating Management" issued by the State Administration of Taxation on July 17, 2003 and the "Tax Credit Management Measures (for Trial Implementation)" promulgated on October 1, 2014. Based on the criteria specified in these documents, the tax credit of a taxpayer within a certain period will be rated as four grades A, B, C, and D, which will be revaluated once in every new tax year. In 2018, the State Administration of Taxation announced "Announcement on the Matters Related to Tax Credit Evaluation", a new M grade was proposed, and the tax credit rating level was changed from A, B, C, and D to A, B, M, C and D. The indicators used to evaluate the tax credit ratings of enterprises are shown in Table 1.2, each indicator has a score, and the total score is 100. The indicators and scores can be decomposed and refined by local governments according to the actual situation. The scores and conditions corresponding to each tax credit rating level are shown in Table A1.

On May 4, 2017, the State Administration of Taxation and the China Banking Regulatory

Commission jointly issued the “Notice on Further Promoting the Work of ‘Interaction between Banks and Tax Authorities’”, which proposed ten measures from four aspects, including improving the cooperation mechanism, increasing the information exchange, promoting the cooperation innovation and strengthening the construction of supporting mechanism. Among the measures, the policies “extend the benefit enterprises from the ones with tax credit rating of A to the ones with rating of B”, “encourage the banking financial institutions to integrate tax data with multi-dimensional enterprise information, research and explore to build an ‘Internet + big data + financial + tax’ platform with internet information technology, and actively innovate credit products” and other policies have received positive responses from the market.

The brand of “Jiangsu Bank” was born during the Revolution of 1911 and has a history of 100 years. In 2007, approved by the former China Banking Regulatory Commission, 10 city commercial banks of Jiangsu Province were combined to form a new Jiangsu Bank. After continuous exploration, Jiangsu Bank has established a transformational policy of “characteristic development, endogenous growth, and innovation-driven”, and formulated a four-in-one strategy of “characterization, intelligence, integration, and internationalization” to promote technology, products, and management. Various reforms and innovations have achieved remarkable results. By the end of 2018, the total assets were 1.93 trillion RMB and the net profit was 13.26 billion RMB. It is the largest corporate bank in Jiangsu Province, listed on the Shanghai Stock Exchange of China on August 2, 2016 (stock code: 600919). It also has been included in the MSCI Index (US Alum Index) and the SSE 50 Index. It ranks 117th (2017) in the British “Banker” magazine’s Tier 1 Capital Ranking of global banks, showing an increase of 243 positions in that ranking from the beginning of the establishment of the bank, and is one of the banks with a faster rise.

Jiangsu Bank launched “Tax E Financing” in 2015, which is the first true online credit loan product of China, creating a financing service model of “online application, automatic approval, real-time arrival, and lending and repaying at any time” for micro and small enterprises.

1.2 Current loan review mechanism of “Tax E Financing”

One-vote veto rule is used in the loan review of “Tax E Financing”. A series of review rules are set by the computer system, and are classified into rejection rules and prompt rules according to the degree of influence. The mark of rejection rule is R, and the marks of prompt rule are H and L. As shown in Table 1.3, which contains 53 rejection rules, the credit application is rejected as long as the enterprise triggers any of them. In addition, there are two L-type and three H-

type rules. If an enterprise triggers the L-type and H-type rules it will not make its credit application be rejected, but the number of those triggers will affect the level of the loan interest rate and the credit line.

Table 1.3 Loan review rule of “Tax E Financing”

Rule code	Rule description
L901	Total overdue times in the previous 24 months is equal to 3.
L902	Total overdue times in the previous 24 months is equal to 2.
R114	Operating entity name contains industry sensitive words.
R115	Business of the operating entity declines seriously.
R116	Tax data is expired.
R117	There is no tax declaration data for the past two months.
R201	The borrower is not the “legal representative” or “responsible person” of the operating entity.
R202	The business license expiration date of the operating entity is less than 12 months from the loan application date.
R203	The borrower’s personal business credit loan amount in our bank is higher than 0. (Including the loan in transit and outstanding).
R204	The borrower’s personal business loan amount in our bank is higher than 5 million RMB (Including the loan in transit and outstanding).
R205	The borrower has an outstanding loan amount of Kayidai in our bank.
R206	The operating entity’s corporate loan in our bank is higher than 0. (Including the loan in transit and outstanding).
R207	Borrower age is lower than 18 years old or higher than 60 years old.
R208	Duration of the operating entity is less than 2 years (calculated by day).
R209	Credit line applied is lower than 1,000 RMB or higher than 1 million RMB.
R210	Credit period applied is shorter than 1 month or longer than 12 months.
R212	The rating of the operating entity does not meet the requirements.
R213A/R213B	The borrower/ operating entity does not meet the requirements in the classification form of our bank.
R214	Borrower is not the sole borrower of the operating entity.
R401/R404	Borrower credit (maximum continuous overdue period of borrower loan or credit card is more than 3 or maximum cumulative overdue period of loan is more than 6).
R407	The bad balance of operating entity is more than 0.
R411	The operating entity credit has overdue records.
R414	The external guarantee claims of operating entity are listed as concerns.
R417	Excessive debt.
R418	High debt ratio.
R421	The borrower’s external guarantee claim is overdue.
R427	The number of queries exceeds limit during the specified time.
R401-R404 R419-R420/R428	The overdue period of borrower loan exceeds limit.
R420	Borrower’s loan/credit card has overdue interest in the most recent period.
R407-R414/R421	The current loan or external guarantee of the borrower or operating entity carries risks or potential risks.
R415/R416/R426	Too many financial institutions grant credit to the operating entity.
R417/R418	The debt tolerance is poor.

R422-R425	Borrower or operating entity has derivative risk in the business.
R501	The borrower is on the blacklist for the write-off.
R502	The borrower is on the blacklist of losses.
R503	The borrower is on the blacklist of escaped debt.
R504	The borrower has executed case that is not completed.
R505	The operating entity is on the blacklist for the write-off.
R506	The operating entity is on the blacklist of losses.
R507	The operating entity is on the blacklist of escaped debt.
R508	The operating entity is on the blacklist of revoking the business license.
R601	Tax payment time is less than 2 years.
R602	Total tax payment <20,000 in the past 12 months.
R603	There is tax owed information.
R604	Tax credit rating of the operating entity is not A or B.
R605	Operating status is not normal.
R606	The total numbers of tax overdue in the previous 24 months are more than 4 times.
R905	For customers with financial statements, the ones that asset-liability ratio is more than 75% will be rejected, and no report does not trigger the rule.
R906	Income tax actually paid in the previous 24 months is less than 0.
R907	Total sales revenue in the previous year is more than 100 million RMB.
R908	The operating entity is a financial, construction or real estate enterprise.
R909	The legal representative of the entity was changed within 90 days.
R910	For customers who pay VAT (value added tax), the total sales income of the last two periods is equal to 0.
R911	The customer has the information of the executed person (still in execution).
R114	The name of the enterprise contains the words “steel”, “steel material”, “stainless steel” and “steel trade”.
H902	The operating has the record of overdue tax.
H904	The investment of borrower accounts for less than 20% of total capital.
H910	VAT has not been declared for three consecutive periods or six cumulative periods.

Source: Jiangsu Bank.

1.3 Research problems

As of the end of 2017, “Tax E Financing” has created considerable economic benefits for Jiangsu Bank, and effectively served the financing needs of micro and small enterprises. A total of 7,242 enterprises have been granted credit by Jiangsu Bank. However, its problems are gradually exposed and mainly reflected in the low approval rate and unreasonable risk assessment.

Taking the data of Jiangsu Bank’s Nantong Branch as an example, Table 1.4 shows the default information of “Tax E Financing” and offline micro and small enterprise loan business of Nantong Branch in the second half of 2017. As can be seen from the table, compared with

the offline business, “Tax E Financing” has higher safety. In the statistical interval, the overdue loan amount of “Tax E Financing” accounted for 1.97% of the loan balance, and the number of defaulted enterprises accounted for 1.48% of the total number of enterprises using loans, which were both lower than 2.28% and 3.83 % of offline micro and small enterprise loan business. At the same time, the credit scale of “Tax E Financing” is much smaller than offline business. In the statistical interval, the loan balance of “Tax E Financing” was 234 million, while the loan balance of offline business was 5.91 billion, which was 25 times than that of “Tax E Financing”. The reason for the low default rate and small scale lies in the extremely low approval rate of “Tax E Financing”.

Table 1.4 Default information of “Tax E Financing” and offline SMEs loan business of Nantong Branch in the second half of 2017

Online			Offline		
Loan balance (100 million)	Default amount (100 million)	Default rate	Loan balance (100 million)	Default amount (100 million)	Default rate
2.34	0.046	1.97%	59.10	1.35	2.28%
Number of enterprises using loans	Number of enterprises defaulted	Default rate	Number of enterprises using loans	Number of enterprises defaulted	Default rate
675	10	1.48%	1698	65	3.83%

Note: Online represents “Tax E Financing”.

Source: Nantong Branch of Jiangsu Bank.

Table 1.5 Approval information of “Tax E Financing” and offline SMEs loan business of Nantong Branch in the second half of 2017

Online			Offline		
Application Amount (100 million)	Approved amount (100 million)	Approval rate	Application Amount (100 million)	Approved amount (100 million)	Approval rate
33.29	8.77	26.34%	39.39	37.16	94.34%
Number of enterprises applying for loans	Number of enterprises approved	Approval rate	Number of enterprises applying for loans	Number of enterprises approved	Approval rate
2087	559	26.78%	1271	1230	96.77%

Note: Online represents “Tax E Financing”.

Source: Nantong Branch of Jiangsu Bank.

Table 1.5 shows the approval information of “Tax E Financing” and offline micro and small enterprise loan business of Nantong Branch of Jiangsu Bank in the second half of 2017. As can be seen from Table 1.5, in the second half of 2017, there were 2,087 enterprises applying for loans through “Tax E Financing”, far more than 1,271 enterprises applying for loans from offline business. However, only 559 of them got the approvals, and the approval rate was only 26.78%, while the approval rate of the offline micro and small business loans was 96.77%. At the same time, the amount of loans applied through “Tax E Financing” was 3.329 billion, of

which only 26.34% were approved, while 94.34% of offline loans were approved.

This thesis calculates the default rate corresponding to the unit approval rate of the loan amount based on the default rate and approval rate of Table 1.4 and Table 1.5. The specific calculation method is that dividing the default rate of “Tax E Financing” of 1.97% by the approval rate of 26.34%. It is found that every 1% of the approval rate corresponds to a default rate of 7.48%. Using the same method to calculate the data of offline loans (dividing 2.28% by 94.34%), it was found that the default rate corresponding to each 1% approval rate is 2.42%. This shows that although the “Tax E Financing” of Nantong Branch has controlled the approval rate of loan amount to a very low level, its default rate has not decreased by the same proportion, showing that the current tax review rules of “Tax E Financing” are too strict and the risk assessment method is not reasonable.

Commercial banks should seek economic benefits as well as social benefits. The economic benefits are mainly reflected in their profit and asset safety, and while the social benefits are mainly reflected in providing credit services and solving the financing difficulties of micro and small enterprises. The achievement of economic and social benefits mainly depends on the bank’s risk control. The loose risk control may lead to good economic benefits in the short run, but the risk exposure will ultimately affect bank’s profit and safety requirements negatively. While if the risk control is too strict, it will seriously affect the scale and social benefits of commercial banks and cannot meet the financial needs of micro and small enterprises. The relevant data of “Tax E Financing” indicates that its risk control is too strict, and the low approval rate is exchanged for the low default rate, at the expense of economic and social benefits in exchange for asset safety, which results in that the total benefits of banks, enterprises and government are not optimal.

The reason why the loan review mechanism of “Tax E Financing” has the above problem mainly lies in the following two points.

First, the tax information of micro and small enterprises is not used sufficiently. As a loan product of “Interaction between Banks and Tax Authorities”, the use of tax information by “Tax E Financing” is mainly reflected in two cases. In the first case, it is used as a threshold for choosing users, meaning that enterprises with tax credit rating of B or above are eligible to apply. In the other case, it is used as a benchmark for determining the credit line. However, in fact, in addition to reflecting whether the company survives and whether it is honest or not, the tax payment amount directly reflects its operating status and profitability, and to a certain extent represents the solvency of the company. Moreover, the data is provided by government authorities and is more authentic and reliable than the data provided by the enterprise itself.

Therefore, tax data should be fully used to assess the default possibility of the company.

Second, the loan review rule of “Tax E Financing” is too rough. The current review rules are too simple, as long as the enterprise triggers any one of the rejection rules, it cannot be approved. This does not provide an overall and comprehensive evaluation of the company’s credit risk, and lacks quantitative analysis to dig deeper into enterprise data. In fact, many statistical methods and machine learning algorithms have been applied in the financial field, and have performed well in credit risk assessment.

1.4 Research questions

The essence of loan review rules is a credit risk assessment model, which is composed of two parts, one is the credit evaluation index system, and the other is the evaluation method. In order to optimize the loan review rules for the “Tax E Financing” of Jiangsu Bank, which means effectively increasing the approval rate under the premise of controlling the default rate, and seeking the synergistic growth of the product’s economic and social benefits, this thesis will study the following two questions.

First, how to build and optimize the credit risk evaluation index system for “Tax E Financing” based on reliable information?

Based on the absolute reliability of tax information, this thesis will refer to the research results of the existing literatures on the credit evaluation index system of small and micro enterprises, and add the tax information of the enterprise into the index system. Combining the data that can be collected by the “Tax E Financing”, this thesis will collect the basic information, financial information and tax information of the sample enterprises, and analyze it from the following two aspects. The first is to profile the defaulting companies and non-defaulting companies based on feature data, and initially study which features can effectively distinguish between defaulting companies and non-defaulting companies, and whether the features of the non-passing companies are similar to the defaulting companies. The second is to test the ability of tax information to predict the default risk. According to the types of information, we combine the basic information, financial information and tax information into different combinations to construct multiple sub-samples. For each sub-sample, support Vector Machine is used to train the classifiers. By comparing the performances of classifiers constructed with different sub-samples, the classification ability of tax information is clarified. The third is based on an effective and reliable feature set, using Random Forest- to further filter and optimize the index system.

Second, how to choose a more effective classification method for “Tax E Financing” to evaluate the credit risk of the enterprise?

Based on the optimized “Tax E Financing” credit risk evaluation index system, this thesis will use machine learning algorithms to train the sample data, compare the performance of different classifiers, and select the best-performing method as the credit risk classification algorithm for “Tax E Financing “. This thesis can collect relevant feature data of defaulting companies, non-defaulting companies, and non-passing companies, The data of defaulting companies and non-defaulting companies belong to labeled samples, and those of non-passing companies are unlabeled samples. Therefore, we can perform supervised learning algorithm based on the labeled data set, such as Support Vector Machine, Neural Network, Logit Regression, Random Forest, and comprehensively use the labeled and unlabeled data set to perform semi-supervised learning algorithm, such as Transductive Support Vector Machine. This thesis will select suitable method from supervised and semi-supervised machine learning algorithms to train and test the data set, and select the best algorithm from them.

1.5 Research structure

This thesis is an attempt to improve the credit risk evaluation mechanism of “Tax E Financing”, to evaluate the credit risk of the applicant enterprise more accurately, and appropriately increase the approval rate to make it closer to offline loan businesses, while maintaining a relatively low loan default rate. The structure of the thesis is as follows:

Chapter 2 will review the relevant literatures from four aspects: the concept and theory foundation of credit risk, enterprise credit risk assessment methods, enterprise credit risk assessment index system, and the information sharing.

Chapter 3 will introduce the research methods of this thesis from three aspects: model evaluation methods, feature processing methods and classification algorithms.

Chapter 4 will introduce the sample data used in this thesis. This chapter will collect the basic information, financial information and tax information of defaulting companies, non-defaulting companies, and non-passing companies. We will profile the default samples to initially explore features that can effectively classify defaulting companies and non-defaulting companies, and analyze the features of the non-passing company to further judge the possibility of misclassification of them.

Chapter 5 will test the ability of tax information to predict the credit risk and further optimize the credit evaluation index system. First, according to the feature type, the collected

sample data is divided into multiple feature subsets, and Support Vector Machine is used to train the classifier. By comparing the performances of the classifiers trained by different subsets, the classification ability of tax information is clarified. Secondly, for the data subset with the best classification performance, the Random Forest algorithm is used to further filter and optimize the index system.

Chapter 6 will use various algorithms of supervised learning and semi-supervised learning to train and test the classifier based on the optimized credit risk evaluation index system. Among them, supervised learning algorithms include individual algorithms and ensemble algorithms. Individual learners include Support Vector Machine, Logistic Regression and BP Neural Networks. In the ensemble algorithm, this thesis mainly uses Random Forest, while in the semi-supervised learning algorithm, Transductive Support Vector Machine is used. This chapter will test whether the optimization of the index system can effectively improve the performance of the model, and choose the best-performing classifier as the credit risk evaluation model of “Tax E Financing”.

Chapter 7 will present the main conclusions of this thesis, and propose references for the loan review mechanism for “Tax E Financing” and the “Bank Tax Interaction” loan products of other banks, and propose corresponding suggestions for the formulation of relevant policies of government departments. Meanwhile, this chapter will point out the research limitations of this thesis.

1.6 Research value and expected contribution

This thesis has two aspects of theoretical values. First, different from existing researches, this thesis adds tax information to the enterprise credit risk evaluation index system. Since the work of “Interaction between Banks and Tax Authorities” has not been carried out for a long-time, Chinese government has not previously opened the tax data of enterprises to banks or other public. Therefore, there are no domestic studies on the credit risk assessment of micro and small enterprises by incorporating tax information into the credit evaluation index system. Foreign studies have also shown that banks in developed countries are more dependent on the information provided by enterprises themselves and the information provided by private credit bureaus when evaluating the credit risk of micro and small enterprises. They rely less on the government to share information, so the use of tax information in their indicator system is also less. By incorporating tax information into the enterprise risk assessment model, the research in this thesis may effectively improve the accuracy of risk assessment, and more accurately

predict the default probability of the enterprise. Second, this thesis compares the performance of supervised learning and semi-supervised learning algorithms. Due to data limitations and other reasons, the existing studies mainly use supervised learning algorithms to train classifiers on labeled data sets when using machine learning algorithms to predict the credit risk of enterprises. However, this thesis can collect more data from companies that have not passed loan approval. Unlabeled data can be added to the sample data, and then semi-supervised learning algorithms can be used to train the classifier. Therefore, the research in this thesis can compare the prediction capabilities of supervised learning and semi-supervised learning algorithms for the credit risk of small, medium and micro enterprises, and provide more empirical evidence for research in related fields.

This thesis has four aspects of practical value. First, through the improving of the loan review mechanism, the interest income of Jiangsu Bank can effectively increase, and more economic benefits can be brought to the bank. Second, by increasing the approval rate and allocating the credit line efficiently, more financial needs of micro and small enterprises can be satisfied, and more social benefits will be gained. Third, through the mechanism improvement, the tax information will play a much more important role in the loan review, which will encourage the micro and small enterprises to pay more attention to their tax paying behavior and create more tax revenue for the government. Fourth, there are more banks joining in the work of “Interaction between Banks and Tax Authorities”, and the research in this thesis can provide reference for them to improve their loan review mechanism.

Chapter 2: Literature Review

2.1 Concept and theoretical basis of credit risk

2.1.1 Credit and credit risk

The concept of credit originated from commercial activities, and was gradually introduced into the social field and the moral category, forming a complex concept with a very wide range of connotations. From the perspective of the object of credit evaluation, it can be divided into categories such as personal credit, corporate credit and government credit. From the perspective of the categories of activities that generate credit, it can be divided into social credit, bank credit, and tax credit (Bernanke & Blinder, 1988). Different disciplines, different industries, and different fields have different understandings and interpretations of credit.

In the economic field, credit is defined as the ability of both parties to settle a contract to perform relevant obligations according to the agreed terms (K. Brown & Moles, 2014). In the financial field, the borrower in the credit contract is more concerned. Credit mainly means that the borrower fulfills the contractual obligations and repays the principal and interest to the lender on time.

Credit risk is defined as the possibility that the parties to a contract will fail to perform their obligations in accordance with the agreed terms. In the banking system, credit risk is further defined as the possibility that the borrower of a credit contract will fail to perform as required. Therefore, credit risk is also known as default risk, performance risk or counterparty risk. K. Brown and Moles (2014) point out that credit risk can be measured by three characteristics. First is the risk exposure. Second is the default probability. And third is the loan recovery rate, that is, how much loan can be recovered if the borrower defaults. In addition to K. Brown and Moles (2014), there are many other works of literature that define default risk. For example, Merton and Bodie (1992) define default risk as a credit default caused by a company suffering losses in economic activities and negative net asset. Stiglitz and Weiss (1981) define and explain default risk from the perspective of information asymmetry. They argue that information asymmetry between borrowers and lenders leads to adverse selection and moral hazard, which in turn lead to borrowers' credit default.

Among the credit of various subjects, corporate credit has received extensive attention from

all walks of life. Because the enterprise sector is one of the main economic sectors of a country, the development level and development quality of enterprises have an important impact on the overall competitiveness of the country. There is a relationship of mutual influence and mutual generation between enterprise development and enterprise credit. On the one hand, the level of corporate credit is determined by the level of development of the company. And on the other hand, the level of corporate credit will in turn affect the level of development of the company (Y. F. Chen, 2020). Corporate credit also includes many aspects, such as the credit shown by the company in production and operation activities, the credit shown in social responsibility, and the credit shown in financial transaction activities. The risk that a company generates in its financial transaction activities is called corporate loan default risk. The corporate credit risk studied in the existing literature, in most cases, refers to the loan default risk.

The impact of credit risk on corporate value has two manifestations. One is the direct economic loss caused by the default of the enterprise, and the other is that the credit risk of the enterprise affects the debt value of the enterprise through various intermediary channels, such as changes in credit level, changes in the ability to perform contracts, thereby affecting the market value of the enterprise, and even leading to business failures (Caouette et al., 1998). In traditional financial activities, more emphasis is placed on the impact of the first level. However, in the modern financial environment, due to changes in the risk environment, and with the improvement and development of risk management technologies and capabilities, more emphasis is placed on the impact of the second level.

The credit risk of a company mainly comes from three aspects. First, the company does not have the ability to perform the contract. For example, due to poor profitability, poor asset liquidity and other reasons, the company cannot repay the principal and interest of the loan on time, resulting in credit default. Or the debt level of corporate is too high, leading to insolvency, default or bankruptcy (Crouhy et al., 2000). Second, the enterprise has the ability to perform the contract, but does not have the willingness to perform. This situation is relatively rare at the enterprise level, and is more common in personal loans or P2P loans (Q. Li, 2018a). The third is corporate defaults caused by changes in the external environment. Corporates have both the ability and willingness to perform the contract, but other factors such as war and political risk lead them to default.

Under the current economic environment of increasing financial innovation, deepening of financial globalization, and gradual relaxation of financial controls, the credit risk of enterprises and individuals is still one of the most important risks of financial institutions. In particular, the credit risk of enterprises has a decisive impact on the healthy development of financial

institutions. In actual activities, banks should consider both the ex-ante expected default risk and the ex-post actual default risk when evaluating and measuring the borrower's default risk. Ex-ante risk refers to the bank's prediction and evaluation of the borrower's default probability based on various indicators and methods (A. N. Berger & Udell, 2002). Meanwhile, banks generally use the non-performing loan ratio to measure their ex-post default risk.

As for how to measure whether the borrower defaults, the definition given by the New Basel Capital Accord is that, when the loan has been overdue for more than 90 days, or the borrower has undergone debt restructuring, the borrower is judged to be in credit default. In China, according to the "Guidelines for the Classification of Loan Risks" issued by the China Banking Regulatory Commission in 2007, corporate loans are classified into five categories: normal, special mention, doubtful, subprime and loss. Among them, the last three categories are defined as non-performing loans. The classification criteria for these five types of loans mainly refer to the number of days overdue. According to the Interim Measures for the Risk Classification of Financial Assets of Commercial Banks (2019), as long as the loan is overdue for more than 90 days, it is defined as a non-performing loan, even if it has sufficient collateral.

2.1.2 Information asymmetry theory

Information asymmetry theory is a classic theory in modern information economics. The basic content of the theory is that, the information held by all parties in the transaction is inconsistent, the party with more and higher quality information will be in an advantageous position in the transaction, and the party with less and lower quality information will be in a disadvantage in the transaction. For example, in a commodity market, commodity sellers have more information about commodity quality, characteristics and prices, while buyers have relatively little commodity information. At this time, commodity sellers will obtain excess profits due to information asymmetry.

Majić et al. (2015) argue that information asymmetry can be divided into ex-ante information asymmetry and ex-post information asymmetry according to the time point when information asymmetry occurs. Ex-ante information asymmetry refers to the difference in information between buyers and sellers before a transaction occurs. At this time, it is generally the seller who has more information about goods and services. Ex-post information asymmetry refers to the information difference between buyers and sellers that occurs after the transaction. At this time, since the goods have been transferred and the services have been provided, generally the buyer has more information. The existence of information asymmetry will lead to

two serious consequences, one is adverse selection, and the other is moral hazard. Among them, adverse selection is mainly caused by ex-ante information asymmetry. Moral hazard is mainly caused by the ex-post information asymmetry.

Adverse selection generally means that when there is information asymmetry between the two sides of the transaction, the buyer with less information gives a lower transaction price for the purpose of protecting his own interests, resulting in a lower market equilibrium price. Adverse selection squeezes quality products out of the market, flooding the market with inferior products, resulting in a decline in overall product quality. Akerlof (1970) builds a model based on incomplete information game theory to discuss the adverse selection problem caused by information asymmetry in the used car market. Akerlof (1970) points out that there is information asymmetry between sellers and buyers of used cars, sellers have more information, and buyers cannot judge the quality difference of each car because they have less information. In this case, in order to avoid losses due to quality uncertainty, the buyer will lower the price. Too low quotations will make high-quality used cars withdraw from the market, flood the market with low-quality used cars, and form a phenomenon of “bad money drives out good money”. This phenomenon is called Gresham’s Law.

The concept of moral hazard originated in the 1980s, and generally refers to the behavior of the transaction party with the information advantage to harm the interests of the other party for the purpose of maximizing his own utility after the transaction is completed. Moral hazard is also called moral crisis. Moral hazard problems are commonly encountered in insurance contracts performance. For example, after purchasing auto insurance, car owners are more likely to make risky driving behaviors, and insurance companies with less information will bear the consequences of risky driving. At the same time, in the principal-agent problem, the agent may harm the principal’s interests due to the information advantage. For example, the management of a company, acting as the agent of shareholders, has more internal information about the company, and shareholders cannot effectively supervise their behavior. At this time, management may hollow out the company’s assets, damage the interests of shareholders, and form moral hazard (Garen, 1994).

Scholars have discussed information asymmetry in various fields, including commodity markets, labor markets, and financial markets. Among them, the information asymmetry in the financial markets, and the resulting adverse selection and moral hazard have received extensive attention. Modigliani and Miller (1959) propose the famous MM theory, which discusses the effect of capital structure on firm value. In a complete capital market, without the influence of taxation, Modigliani and Miller (1959) find that capital structure has no effect on firm value.

However, Stigler (1961) finds that the acquisition of information is not cost-free due to the existence of information asymmetry. The existence of information acquisition cost and search cost leads to the incompleteness of the market. In an incomplete market, information asymmetry will make enterprises face financing constraints, and the degree of financing constraints is significantly positively related to the degree of information asymmetry faced by enterprises.

Bollazzi et al. (2019) study the impact of information asymmetry on the deal selection process of Italian venture capital. They argue that there is a huge information asymmetry between venture capitalists and start-ups seeking venture capital, and that venture capitalists have a severe lack of understanding of startups. Bollazzi et al. (2019) conduct an empirical study using data from the Bureau van Dijk database and a survey of Italian venture capital funds. They find that the educational level of the start-up management team, international management experience, and the number of patents is significantly positively related to their probability of obtaining venture capital. Their research shows that in the venture capital market, start-ups should transmit relevant information as much as possible to reduce the degree of information asymmetry and avoid adverse selection.

The credit market is an important component of the financial market. There is also information asymmetry, which has an important impact on credit transactions. In a lending relationship, the borrower is the buyer and the lender is the seller. Before the lending transaction, the borrower has an information advantage, and the lender has an information disadvantage. When the borrower is a large enterprise, especially a listed enterprise, it discloses more information publicly, which can help reduce the information asymmetry between the borrower and lender to a large extent, and help banks and other financial institutions to correctly assess the default risk of the enterprise. Therefore, it is easier for them to obtain debt financing.

However, for SMEs or individual financiers, their public disclosure information is less, and the standardization, objectivity and integrity of the information are difficult to guarantee. This will cause serious information asymmetry between them and financial institutions and lenders of P2P platforms. At the same time, since their own repayment ability is not as high as that of large enterprises, and credit risk is difficult to assess, this will lead to adverse selection of financial institutions and P2P lender. Therefore, in the case of information asymmetry, it is difficult for many SMEs and individuals to obtain debt financing. Even if they obtain loans, the interest rate is very high. In fact, individuals and businesses willing to obtain loans at high interest rates generally have a higher risk of default. This leads to the “bad money drives out good money” in the credit market, borrowers with good credit cannot get loans at a reasonable price, and borrowers with poor credit get loans, but they also have extremely high moral hazard.

Information asymmetry in the credit market leads to credit rationing policies. Credit rationing refers to the behavior of banks and other financial institutions to control the scale of loan issuance. Jaffee and Russell (1976) study the credit market with asymmetric information and find that credit rationing can effectively solve the adverse selection of financial institutions. Stiglitz and Weiss (1981) also believe that in a credit market with asymmetric information, credit rationing by financial institutions is a rational choice, which can effectively reduce adverse selection and moral hazard. Williamson (1987) discusses the optimal contractual arrangement between lenders and borrowers in a credit market with asymmetric information and costly monitoring costs by establishing an equilibrium model. The study finds that credit rationing is an equilibrium choice, which is consistent with Stiglitz and Weiss (1981).

In order to reduce the information asymmetry in the credit market, the role of financial intermediaries has been emphasized. Financial intermediaries can provide lenders with supplemental information about borrowers to help them conduct credit risk assessments and avoid adverse selection. Padilla and Pagano (2000) study the effect of credit bureaus on adverse selection of banks. They find that when lenders shared information, credit rationing is reduced and credit scale increased.

Market signals can also eliminate the effects of information asymmetry to a certain extent. When Spence (1973) studies the information asymmetry in the job market, the study finds that employers can use various market signals to assess the productivity of job seekers, and market signals can avoid employers' adverse selection. But at the same time, the study also emphasizes that market signals can effectively transmit information about job seekers only when there is a significant difference in signaling costs.

2.2 Research on enterprise credit risk assessment methods

The development of enterprise credit risk assessment methods can be divided into two stages. The first is the qualitative credit risk assessment stage, and the second is the quantitative evaluation stage brought by the development of econometrics, modern financial engineering and artificial intelligence.

2.2.1 Qualitative credit risk assessment methods

Before the 1950s, the expert evaluation method was widely used in assessing the credit risk of enterprises. Experts scored the credit risk of each enterprise according to a certain index system, such as 5C, 4F, 5P, 5W. The earliest credit evaluation index system was the 4C standard

proposed by American banker William Post in 1910, which mainly included four aspects: character, capacity (solvency), capital, and collateral. On the basis of 4C, banker Edward F. Gee added the “Condition of business”, forming the 5C standard. Thereafter, in practical applications, 5C was summarized into 3 factors, named 3F. Among them, the character and capacity are summarized as personal factors, the capital and collateral are summarized as financial factors, and the condition of business is summarized as economic factors. Later, organizational factors were added, and then 3F was developed into 4F. 5P refers to the five dimensions of person, purpose of capital, payment source, protection of the debt, and perspective. 5W refers to the five dimensions of the borrower (Who?), the use of the loan (Why?), the term of the repayment (When?), the collateral (What?) and how to repay (How?). The expert evaluation method has obvious defects, such as rough method indicators, subjective standards, difficulty to unify, and its results are not stable.

2.2.2 Quantitative credit risk assessment methods

2.2.2.1 Multivariate credit risk assessment methods

The multivariate credit risk assessment method mainly uses the financial characteristics of the enterprise as explanatory variables to explain and predict the probability of corporate bankruptcy, financial crisis or debt default. This method is widely used by academia and industry, and mainly includes two major categories. The first category is a credit risk assessment method based on a multivariate discriminant model, and the second category is a credit risk assessment method based on regression models, such as the Logit and Probit models.

(1) Credit risk assessment method based on multivariate discriminant model

The earliest quantitative credit risk assessment method is the linear discriminant method proposed by Altman (1968). This takes corporate bankruptcy as a measure of financial crisis, following the approach of Fitzpatrick (1932). It selects 33 bankrupt companies and 33 non-bankrupt companies to form a paired sample, selecting five core explanatory variables, such as retained earnings/total assets and working capital/total assets from 22 financial ratios, and constructs the Z-score model. Through their weighted average, these five financial ratios are combined into a comprehensive score, and the Z value of each enterprise is then obtained. The credit risk of the enterprise is judged according to the interval in which the Z value is located. Research shows that when the Z value is greater than 2.675, the credit risk of the enterprise is small. When the Z value is less than 1.8, the enterprise faces greater and significant credit risk. However, there are three shortcomings in this study. First, when the Z value is between 1.8 and

2.675, the credit risk of the enterprise cannot be judged. Second, the accuracy of the model decreases year by year. The model has a predictive accuracy of 95% in the year before the bankruptcy, and 72% in the two years before bankruptcy. Third, the sample data are all from manufacturing industries, and there may be industry limitations, so that the model may not be applicable to other industries.

The Z-score model is extended by Deakin (1972). This study uses the financial indicators of Beaver (1966) to establish a quadratic function, and constructs a discriminant model for each year's data, instead of establishing a discriminant model for the whole sample as in Altman (1968), which significantly improves the prediction accuracy. What needs to be added is that Beaver (1966) conducts a comparative study of companies that have experienced financial crisis and those which have not yet experienced financial crisis, and finds that the former's (i.e. companies that have experienced financial crisis) cash-to-debt ratio, return on assets, and current ratio are significantly lower than the latter's (i.e. companies which have not yet experienced financial crisis).

Thereafter, in order to further improve the above problems of Z-score model, Altman et al. (1977) proposes the ZETA model. The model uses seven financial indicators (including return on assets) as explanatory variables, extends its application field to other industries such as retail, and has higher accuracy, becoming the mainstream analysis tool in the field of credit analysis. A large number of literatures use similar discriminant method models to assess the bankruptcy risk of enterprises, such as Blum (1974), Wilcox (1971, 1973) and Deakin (1977).

There are also a lot of literatures applying multivariate discrimination method to assess the credit risk of certain industries or certain types of enterprises. For example: Edmister (1972) applies it to small businesses; Altman (1973) applies it to the rail industry; Altman and Loris (1976) apply it to over-the-counter brokers; Edmister and Schlarbaum (1974), Santomero and Vinso (1977) and Pettway and Sinkey (1980) apply it to the banking industry; Schipper (1977) applies it to the education industry; and finally, Pinches and Trieschmann (1974) apply it to insurance industry. These studies use different indicators and different data to construct new models, expanding the application scope of multivariate discriminant models from large enterprises to medium-sized and small-sized ones, from manufacturing industries to non-manufacturing ones.

There are also literatures that use discriminant methods to assess the credit risk of Chinese companies. For example, S. Lu and Ooyang (1999) collect financial data of Chinese listed companies from 1992 to 1995 and evaluate the credit risk of all 296 companies based on the Z-score model proposed by Altman (1968).

(2) Credit risk assessment method based on multivariate regression models

As the multivariate linear discriminant model has high predictive accuracy it is still often used by many banks, but it also has its limitations. First, there is no uniform standard for the selection of explanatory variables. Different indicators may significantly affect the discriminant results of the model. Second, the model assumes that there is a linear relationship between variables; therefore, it cannot deal neither with nonlinear relationships nor with qualitative variables. Third, the model has strong premise assumptions. For example, the independent variables of the model must satisfy the assumption of normal distribution, and the covariance of the bankrupt sample and the non-bankrupt sample must be equal.

Ohlson (1980) argues that the premise assumptions of the discriminant model are difficult to satisfy. For example, the covariance of bankrupt companies and non-bankrupt companies is very different, and usually financial ratios do not meet the assumption of normal distribution. This leads to a decrease in the effectiveness of predictions made using the discriminant method. More importantly, the scores calculated with the discriminant method may not have intuitive interpretation or economic meaning. Therefore, Ohlson (1980) proposes to use the Logit model to predict the financial risk of the enterprises. The Logit model is based on the cumulative probability function, does not need the independent variable to satisfy the normal distribution hypothesis, nor the two groups of samples with the same covariance, and the predicted score of the model is the conditional probability of enterprise bankruptcy, which has a clear economic meaning. Under normal circumstances, 0.5 is used as the threshold. When the bankrupt probability is greater than 0.5, the company is considered to have greater and significant financial risk. When the bankruptcy probability is less than 0.5, the enterprise is judged to be in a financially normal state. Ohlson (1980) selects 2058 non-bankrupt companies and 105 bankrupt companies based on public financial statement information, uses Logit model to predict the financial risk of the companies, and finds that the Logit model has higher prediction accuracy than the discriminant model.

Since then, Logit Regression (LR), as an effective and robust prediction method, has been widely used in credit risk assessment. For example, Maddala (1983) uses it to evaluate the credit risk of private borrowers. The results show that 0.551 as the critical value can more effectively distinguish whether personal loans default. Charitou et al. (2004) use the Logit method to predict the risk of corporate failure for 51 pairs of failed and non-failed UK public industrial companies with the data from 1988 to 1997. The study finds that the company's cash flow, profitability and financial leverage have predictive capabilities for an enterprise's failure risk. The prediction accuracy of this model in the year before the failure of the enterprise is 83%.

There are similar researches such as Gentry et al. (1984), Keasey and Watson (1987), Aziz et al. (1988), Platt and Platt (1990), Ooghe et al. (1995), Mossman et al. (1998), Charitou et al. (2004).

The Probit model is similar to the Logit model, but its assumptions are slightly different. The Probit model assumes that the company's default probability follows a normal distribution. For example, the Riskcalc model used by Moody to evaluate the credit default probability of unlisted companies is based on the Probit model (Dwyer et al., 2004). Zmijewski (1984) uses the Probit model to predict the bankruptcy probability of listed companies in the United States from 1972 to 1978, mainly considering the influence of sample selection bias on prediction accuracy. The study finds that sample selection bias has a significant impact on intra-group discrimination and prediction rate. However, it does not significantly affect the statistical parameters and the overall prediction accuracy of the model.

Unlike the discriminant method, which mainly uses financial indicators of the enterprise and is therefore not suitable for SMEs, many literatures use the Logit / Probit model to conduct empirical research on the credit default risk of SMEs in different countries. Altman and Sabato (2007) use panel data from more than 2,000 US SMEs (sales less than \$65 million) between 1994 and 2002, applying the Logit regression model to predict their credit risks, and find that the accuracy of Logit model is 30% higher than the general model. Behr and Guettler (2007) use the Logit model to predict the loan default probability of German SMEs. Fantazzini and Figini (2009a) compare the prediction effectiveness of a series of classical models and Bayesian longitudinal models on the default probability of SMEs. The study uses panel data from German SMEs to evaluate these models by comparing their intra- and extra-sample prediction capabilities. The results show that the Bayesian model is better than the classical longitudinal models and the mixed Logit model for the intra-sample prediction. For the extra-sample prediction, there is no statistically significant difference between the Bayesian model and the mixed Logit model. Pederzoli and Torricelli (2010) use the Logit model to evaluate the credit risk of SMEs in the Emilia Romagna region of Italy. The study finds that only equity/total assets, EBIT/total assets, sales/total assets and long-term liabilities/total assets have significant explanatory powers for default risk. Pederzoli et al. (2013) have conducted similar research. Fidrmuc and Hainz (2009) collect the loan data of 700 Slovakia SMEs from 2000 to 2005, using the Probit model to study the default probability of SMEs, and find that liquidity and profitability are the decisive factors affecting default risk, and the amount of debt also significantly increases the probability of default.

There are also many studies that assess the credit risk of Chinese companies and individuals

with the Logit or Probit model. For example, Fu and Li (2005) evaluate the credit risk of Chinese listed companies with their annual report data by using the Logit model, and find that the interest coverage ratio and inventory turnover rate are key factors affecting credit risk. Dong (2011) uses the Logit model to assess the credit risk of micro-enterprises. Although the indicators selection and evaluation process are detailed, the overall number of indicators is still too small, and the model prediction accuracy needs to be improved. Fang et al. (2014) estimate the default risk of Chinese individuals with their default data of credit card, using the Lasso-Logistic model, and find that Lasso can quickly and accurately select the key factors affecting consumer credit risk. Compared with the full-variable Logistic model and the stepwise regression logistic model, the Lasso-Logistic model has higher prediction accuracy.

2.2.2.2 Credit risk assessment method based on modern financial engineering tools

The above-mentioned credit risk assessment methods based on the multivariate discriminant model and multivariate Logit and Probit models mainly depend on the financial information of the enterprise, so there are two problems. First, financial information generally only reflects the book value of accounting items such as assets and liabilities, and therefore may not reflect their market value. Second, financial information reflects the operating conditions and financial risk of the enterprise in the past, and cannot reflect the current or future financial risk of the business.

In this context, the industry has gradually developed a credit risk assessment model based on the market value of corporate assets and liabilities, and has broken through the assumption about the linear relationship between corporate financial risk and financial ratios in the aforementioned two multivariate credit risk assessment models, such as the KMV model created by KMV in 1995 (Crosbie & Bohn, 2019) and later extended by Longstaff and Schwartz (1995). KMV model is based on the option pricing theory proposed by Scholes and Black (1973) and Merton (1974). Equity is regarded as a call option held by shareholders. When the value of the enterprise exceeds the liability, the value of the option is positive. When the value of the enterprise is lower than the liability, the value of the option is 0, and the enterprise will have debt defaults. By calculating the probability distribution of the enterprise value, the model can derive the probability of the company's debt default and the expected default loss. The model has a strong theoretical foundation and reduces the dependence on historical data, but it is mainly applicable to listed companies, because the value of an enterprise depends mainly on the market value of its shares. At the same time, the model also has strong assumptions, including the enterprise value follows a normal distribution and the firm's capital structure remains unchanged.

In 1997, J.P. Morgan and KMV jointly launched the Credit Metrics model (Gupton et al., 1997). Based on the VaR (Value at Risk) theory, the model determines the ending-of-period distribution of the market value of the credit assets or credit asset portfolio held by the enterprise according to the credit rating and transfer probability, and calculates the minimum market value under a certain confidence level. Nyfeler (2000) and Jones and Mingo (1999) conduct a detailed analysis and in-depth development of this model.

In addition to the above two models, there are two representative quantitative credit risk assessment models in the world today. They are the Credit Risk + model developed by Credit Suisse First Boston Bank in 1997 (Gordy, 2002) and the Credit Portfolio View (CPV) model proposed by McKinsey in 1998. The Credit Risk + model assumes that the number of defaults follows the Poisson distribution, and calculates the default probability of the company by estimating the parameters of the Poisson distribution. The above four models have different definitions of credit losses. The KMV model and the Credit Metrics model are all mark-to-market models, which are based on changes in the market value of loans. The Credit Risk + is a default model, which focuses on measuring default losses. The Credit Portfolio View model has both properties.

Some scholars have discussed the applicability of the above models in China. For example, N. Fan (2002) proposes that, based on the imperfect credit system in China and the incomplete historical data of enterprises, the models relying heavily on complete historical information are still difficult to be implemented. Therefore, this study introduces the Credit Metrics models. There are also literatures that apply the above models to the credit risk assessment practices of Chinese companies. L. Zhang et al. (2004) evaluate the credit risk of Chinese listed companies with the KMV model for the first time. The research results show that the KMV model can identify the credit risk of enterprises two years in advance and identify the trend of credit risk four years in advance. Ma (2006) finds that the KMV model has predictive power for the financial distress of Chinese listed companies. Zai et al. (2007) select 15 Special Treatment (ST) and 15 non-ST Chinese listed companies, collecting their data in 2005 for empirical research, and find that KMV model can effectively identify ST and non-ST companies. Huo et al. (2011) and S. Zeng and Wang (2013) apply the KMV method to the credit risk prediction of agricultural and manufacturing listed companies. A similar study is done by S. Chen et al. (2010).

Since then, there have been many literatures on improving the traditional KMV model. For example, N. Zhang and Zhang (2010) revise the default point parameters in the traditional KMV model, collecting the data of 82 listed companies to calculate their default distance, and find that the revised model is more suitable to reflect the credit risk situation of Chinese enterprises.

J. Zhou (2009) revises the KMV model based on the GARCH (1, 1) model and finds that the modified model is more suitable for the Chinese market. K Yang (2015) is similar to J. Zhou (2009). K Yang (2015) derives the expected growth rate from the price-earnings ratio, introducing it into the KMV model, and predicts the volatility of the enterprise value through the EGARCH (1, 1) model. This study collects data from 13 Growth Enterprises Market (GEM) listed companies and 13 SME board listed companies, and conducts a credit risk evaluation. It is found that the revised KMV model can better identify the difference in credit risk between the two types of enterprises and the trend of change in credit quality. Dai et al. (2016) propose a KMV model based on historical financial information and the company's newly added debt data. Through the research on listed companies, it is found that this method can improve the prediction accuracy of the KMV model.

2.2.2.3 Credit risk assessment method based on artificial intelligence

With the continuous development of data mining technology and machine learning, more and more literature focus on the use of new models or data mining tools to improve the prediction accuracy of corporate credit risk. The methods used in the existing literatures mainly include three categories: artificial neural network (ANN) analysis, Random Forest (RF) method and support vector machine (SVM) method.

(1) Artificial Neural Network analysis (ANN)

The ANN analysis is the earliest artificial intelligence method used in the assessment of enterprise credit risk. It is derived from the cross-fusion of neuropsychology and applied mathematics. It has no strict assumption about the sample data, can effectively deal with nonlinear problems and non-normal distribution data, and the result obtained is the credit default probability of the enterprise, which has good economic implications. Odom and Sharda (1990) first apply ANN to corporate credit risk assessment, comparing it with traditional multivariate discriminant methods, and find that the accuracy of ANN is higher than discriminant analysis. Altman et al. (1994) also find that ANN has higher prediction accuracy than multivariate discriminant, and Lee and Chen (2005) find that ANN is also superior to Logit model. Trippi and Turban (1996) apply ANN to risk assessment of consumer credit, household mortgages, and bank credit.

ANN is widely used in credit risk prediction of SMEs. Salchenberger et al. (1992) use ANN to classify and predict the financial status of SMEs and compare them with traditional statistical models. The study finds that if the same financial data is inputted for all models, the ANN method requires fewer assumptions, but can achieve higher prediction accuracy and is more

robust.

Sharda and Wilson (1996) analyze the problem of eliminating performance measurements and conduct an ANN experiment in business failure prediction. Although the ANN method has stronger credit risk prediction ability than Logit regression and multivariate discriminant analysis, Lee and Chen (2005), Chung and Gray (1999) and Craven and Shavlik (1997) have shown that the ANN method also has some limitations, such as the long training time, which limits its applicability in dealing with credit risk prediction problems. Therefore, some scholars have begun to combine ANN with Logit regression, to improve both the prediction accuracy of the model and its operational efficiency. For example, S. L. Lin (2009) and Falavigna (2006) combine the Logit model with ANN to form a two-stage hybrid model and explore whether the two-stage hybrid model is superior to the traditional LR and ANN.

There are also literatures that use the ANN to predict the credit risk of Chinese SMEs. For example, C. Wang et al. (1999) take the short-term debt default data of 95 companies as a sample, and find that ANN has better robustness than discriminant analysis. Deng et al. (2010) and S. Z. Bai and Li (2013) also demonstrate that ANN is more suitable for predicting the credit risk of SMEs in the current stage of the Chinese credit market. However, Xiong et al. (2009), S. B. Bai (2010) and Bei et al. (2012) argue that the Logit model is more suitable than the ANN model for predicting the credit risk of Chinese SMEs. Zhu et al. (2016) combine Logit regression and ANN to construct three types of two-stage hybrid models, to predict the credit risk of Chinese SMEs in supply chain financing. The sample used is the quarterly financial information and other non-financial data for 77 listed SMEs and 11 listed core companies from 2012 to 2013. The research results show that the third type of two-stage hybrid model has the best classification ability to predict the credit risk of SMEs in supply chain finance.

Qian et al. (2013) combine the Principal Component Analysis (PCA) and Back Propagation (BP) neural network to evaluate the credit risk of micro and small enterprises. The PCA is used to reduce the dimensionality of multiple indicators. They collect information of 101 credit customers of a bank in Jiangsu Province, 30 of which are used as experimental samples, and the remaining 71 are used as test samples. The prediction accuracy of the method is 85.78%, which has a good fitting effect. Wen and Han (2017) use Multi-layer Perceptron (MLP) neural network to identify the credit risk of technology-based SMEs. Similar studies have been conducted by Guo et al. (2013), H Zhang and Li (2017), Hu et al. (2012), and Meng et al. (2018).

(2) Random Forest method (RF)

Breiman (2001) first proposes the Random Forest method. It is a combined classifier algorithm, and the basic unit of the method is the decision tree. It integrates multiple decision

trees through the idea of integrated learning. The method is widely used in various fields such as marketing, health care insurance and disease risk prediction. Umezawa and Mori (2008) apply it to the credit risk assessment of power companies firstly. Kartasheva and Traskin (2011) find that the Random Forest method can effectively identify the solvency of various insurance companies. Tanak et al. (2016) find that the Random Forest model has significant early warning capabilities for banking crises in Organization of Economic Cooperation and Development countries. However, Fantazzini and Figini (2009b) compare random survival forests model and Logit regression model with data from 1003 small business in the German rating agencies from 1996 to 2004, and find that the Logit model has higher prediction accuracy than the Random Forest model. Y. Wang et al. (2020) compare the performance of five widely used classifiers, including naive Bayes models, Logistic Regression, Random Forest, Decision Tree, and K-nearest neighbor. The study finds that Random Forest performed better in precision, recall and AUC.

In China, C. Lin and Peng (2007) use the Random Forest analysis for the first time to select indicators in corporate credit assessment to improve the prediction accuracy. Fang et al. (2010) apply the improved nonparametric Random Forest classification method to the risk assessment of credit card users and find that the method is superior to the Logit model and the SVC model. Meng (2014) uses the Random Forest algorithm to predict the financial failure of Chinese listed companies and compares it with other methods. The study finds that the Random Forest model has higher prediction accuracy and better robustness than the Logit model, SVM model and ANN model. Y. Zhao (2015) uses the Random Forest algorithm to evaluate the credit risk of the financing platform of Chinese local government. Sheng et al. (2016) use five models to forecast the credit risk with the credit rating data of 330 Chinese listed companies from 2003 to 2015. The empirical results show that the Random Forest classification model has a stronger ability to predict the changes in credit ratings. Rao et al. (2020) build a two-stage Syncretic Cost-sensitive Random Forest (SCSRF) model to assess the credit risk of Chinese rural borrowers. The study finds that the classification ability of SCSRF exceeds that of Random Forest, Logit regression, Classification and Regression Trees and Cost-sensitive Random Forest models.

There are also literatures, which combine the Random Forest method with other methods to improve the prediction accuracy. Prinzie and Van den Poel (2008) combine the Random Forest method with the traditional Logit model, which improves the accuracy of credit risk prediction. Yeh et al. (2012) combine the KMV model with the Random Forest model, calculating the default distance according to the KMV model first, and then input the default

distance into the Random Forest model. This study finds that the hybrid model has better classification effect. Yeh et al. (2014) combine the Random Forest model and the rough set method to construct a hybrid model.

(3) Support Vector Machine method (SVM)

Support Vector Machine is a type of generalized linear classifier that performs binary classification of data by supervised learning, and its decision boundary is the maximum margin hyper plane solved from the learning samples (H. Li, 2012; Vapnik, 1998; Z. Zhou, 2016). This method can use kernel functions for nonlinear classification (Hsieh, 2009), has the advantage of small samples (X et al., 2009), and is widely used in portrait recognition, text classification and other fields (Qin & He, 2005; A. Sun et al., 2002).

A. Fan and Palaniswami (2000) use SVM to predict the financial distress of Chinese companies for the first time. Since then, a large number of scholars have conducted researches and expanded this method. Min and Lee (2005) use grid search technology to find the optimal parameter value of the SVM kernel function by using 5 times cross-validation. The results find that the SVM method is more accurate than multivariate discriminant method, Logit model and BP neural network. Q. Zhang et al. (2015) combine SVM with Logistic regression, Duan and Lu (2015) construct a cost-sensitive SVM model, and Han and Han (2017) apply the orthogonal SVM to enterprise credit evaluation. Y. Xu and Pan (2016) combine Analytic Hierarchy Process (AHP) and SVM to assess personal credit risk. They use AHP to calculate the index weights, process the data according to the weight, and then use SVM to classify the data. The study finds that using AHP to process the data can effectively improve the prediction accuracy. H. Zhang et al. (2021) propose a Firefly Algorithm Support Vector Machine (FA-SVM) for supply chain financial credit risk assessment. Through application analysis, it is found that FA-SVM outperforms LIBSVM.

Meanwhile, some scholars have studied the SVM integration algorithm. Artificial intelligence algorithms are divided into integrated algorithms, combined algorithms and single algorithms. Xia (2015) believes that integrated algorithms and combined algorithms have higher prediction accuracy than single algorithms, but Tsai and Wu (2008) find that the precision of single neural network algorithms is higher than integrated t neural networks. The main reason is that neural network algorithms rely on large samples, while the current loan data collected by commercial banks is more limited. The integration algorithm requires that the total sample be divided into small samples, and the reduction of single training sample size will reduce the prediction accuracy of the model. In this context, some scholars begin to study SVM integration algorithms. For example, Xiao et al. (2016) study the integration of SVM and

statistical methods. J. Chen et al. (2016) study the integration of multiple sub-Support Vector Machines. B. Li et al. (2006), J. Tian et al. (2011) and Lo et al. (2015) find that the SVM integration method has higher prediction accuracy than a single SVM.

SVM is widely used in the credit risk assessment of SMEs because of its advantages in dealing with unbalanced and small sample data and nonlinear and high dimensional relationships. There are many related studies, such as H. Zeng et al. (2013), X. Xu et al. (2014), S. Li et al. (2018), Meng et al. (2018) and Xia (2019).

(4) Other machine learning methods

In addition to Random Forest, there are many literatures that study the application of other tree algorithms in credit evaluation, such as Decision Tree (DT), Regression Tree (RT), Gradient Boosting Decision Tree (GBDT), Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGboost). For example, Alzamora et al. (2022) use the LightGBM model to assess the risk of microfinance in rural Peru. The study finds that this model increases the loan success rate to 96.2%. Z. Tian et al. (2020) use GBDT for credit risk evaluation and find that its accuracy, F1 score and AUC value are better than methods such as SVM, DT and MLP. B. Yu et al. (2022) evaluate and compare the capabilities of multiple machine learning models using credit rating data from 2010-2019 for 355 environmental companies in euro area. The study finds that RT has the highest accuracy, and the classification effect of the mixed use of RF and RT is better.

Some literature studies the application of fuzzy theory in credit evaluation. Cao and Zhang (2022) use the Fermatean fuzzy TOPSIS method to assess the credit risk of 10 Chinese quantum communication listed companies. It is found that Fermatean fuzzy TOPSIS has higher predictive power compared to traditional TOPSIS. Y. Sun et al. (2022) propose a fuzzy decision-making method to assess the credit risk of enterprises, and use the data of 1820 Chinese small industrial enterprises to test the model's classification performance. It is found that the classification performance of LR, k-Nearest Neighbors (k-NN), SVM, Naive Bayes and RF all become better after data preprocessing by Triangular Fuzzy Numbers compared to data processing by expert Grading Method.

There are also literatures discussing credit classification models for imbalanced data, and multi-stage machine learning algorithms for classification. For imbalanced data, Y.-C. Chang et al. (2018) adopts the multi-stage classifier XGBoost to build a credit risk assessment model for financial institutions. They find that XGBoost has higher ROC area and classification accuracy than single-stage classifiers such as Logit model and SVM. K. Wang et al. (2022) construct a two-stage credit evaluation model, using the Logistic model for feature selection,

and then using XGBoost to train the classification model. Fonseca et al. (2020) construct a two-stage credit evaluation model combining fuzzy reasoning and ANN to evaluate the customer credit of private credit card providers in Brazil. Hou et al. (2020) develop a novel combinatorial Dynamic Ensemble Selection (DES) model for imbalanced datasets. When evaluating the performance of the classifier, they use Dynamic Ensemble Selection Multi-Class Imbalance to weight the minority class. The meta-learning framework of META-DES and DES-KNN are comprehensively used to ensure the classification performance and diversity of dynamic ensemble classifiers. Through empirical research on P2P loan data, it is found that the classification accuracy of DES is higher than that of XGBoost and LightGBM. To address the problem of imbalanced data, G. Yao et al. (2022) propose an SVM Ensemble Model with Artificial Imbalance Rate (SVME-AIR), and a new ensemble Feature Selection-Multiple Ranking Information (FS-MRI) model. A hybrid model is formed by combining SVME-AIR and FS-MRI through a two-stage model. They use the supply chain data of Chinese listed companies to conduct empirical tests. It is found that FS-MRI is the optimal feature selection method, and SVME-AIR has the best classification performance.

Several literatures discuss the application of unsupervised machine learning methods in credit risk assessment. For example, Lappas and Yannacopoulos (2021) propose an enterprise credit evaluation model that considers expert opinion as input. Most credit evaluation models are automatic evaluations of various features. As stated before Lappas and Yannacopoulos (2021) propose that expert opinions should be added to the credit evaluation process, allowing experts to create clustering scenarios in unsupervised machine learning algorithms. Liu et al. (2019) improve traditional clustering methods and propose a new Evolutionary Multi-Objective Soft Subspace Clustering algorithm for credit risk assessment.

There are also literatures that mix supervised machine learning and unsupervised machine learning to assess credit risk. Machado and Karray (2022) mix different supervised and unsupervised learning methods and compare the performance of each mixed model. The unsupervised learning method they use is the k-means clustering method, and the supervised learning method includes six algorithms: Adaboost, GB, DT, RF, SVM and ANN. Using data from a North American bank to train corresponding classification models, they find the following results. First, DT, RF, and GB perform the best classification when using a single supervised learning algorithm for classifier training. Second, after each supervised learning algorithm and unsupervised learning algorithm k-means are mixed, the classification performance is improved. Third, among all single algorithm classifiers and hybrid algorithm classification, k-means + RF has the best classification performance. T. Wang et al. (2022)

propose a new multi-classification model of credit risk based on information fusion theory, which combines six algorithms of DT, RF, AVM, k-NN, BP neural network and XGboost. Compared with a single classification model, this fusion classification model has higher accuracy and wider application range.

There are also literatures discussing the problem that it is difficult to establish a reliable machine learning model with small samples. Yu and Zhang (2021) propose a Virtual Sample Generation (VSG) method that integrates Multi-Distribution General Trend Diffusion and Particle Swarm Optimization algorithms for small sample online loan credit risk assessment. This method can expand the number of samples and improve the classification performance of the model. L. Yu et al. (2022) propose a VSG method with feature engineering based on Extreme Learning Machine for the problem of low-dimensional small samples. Among them, VSG is used for sample expansion, and feature engineering is used to improve the data dimension. X. Zhang et al. (2022) propose a data augmentation and hybrid feature selection method based on Wasserstein Generative Adversarial Networks (WGAN) for the high-dimensional small sample. They use WGAN to expand the sample size, and then use a Kernel Partial Least Squares Quantum Swarm Optimization algorithm to solve high-dimensional problem. For the problem of missing data in credit evaluation, F. Zhao et al. (2022) propose a new imputation method called Multiple Generative Adversarial Imputation Network.

2.3 Research on credit risk assessment index system

There are two types of information in financial markets: hard information and soft information. They are often used by banks to assess the credit risk of businesses or borrowers. Liberti and Petersen (2019) define hard and soft information. They consider hard information to be quantitative, easily stored, and disseminated in an impersonal form. While, soft information is difficult to quantify, needs to be understood in a specific context, and has value in a specific environment. Hard information mainly includes financial information provided by enterprises or individuals themselves, as well as financial behavior information retained by enterprises or individuals in banks, all of which are structured and standardized. Soft information is composed of personalized non-structural information, mainly including non-financial information of enterprises or individuals. Compared with soft information, hard information has higher transfer efficiency and processing efficiency, but also has higher information loss (Liberti & Petersen, 2019). Therefore, soft information is an effective supplement to hard information, especially when hard information is insufficient and unreliable.

Although there are obvious differences between hard information and soft information, some soft information can also become hard information after being quantified. For example, when rating agencies conduct credit rating on an enterprise, they consider not only the hard information such as the financial status, but also the non-financial, and then comprehensively evaluate the credit rating of the enterprise. Therefore, the credit rating is a hard information, but it is quantified from soft information. Banks mainly use hard information when assessing the credit risk of large companies, especially listed ones. For the credit risk assessment of small, medium and micro enterprises and individuals, they rely highly on various soft information. B. Zhou (2021) compares and studies the predictive ability of hard information and soft information on default behavior when conducting credit evaluation on Internet financing of SMEs based on BP neural network. The study divides the feature set into three groups: hard information, soft information, hard information + soft information, and input them into the BP neural network for classifier training. The results show that the classification effect of hard information + soft information is the best, the classification effect of hard information is second, and the worst is to use only soft information. This shows that on the premise of using hard information, adding soft information can effectively improve the accuracy of default risk assessment and reduce the loan risk of financial institutions.

2.3.1 The application of hard information in credit evaluation

Hard information mainly includes two aspects, one is the business activities information of the enterprise, and the other is the financial behavior information of the enterprise and of its owner. The following point will summarize the existing research from these two aspects.

2.3.1.1 Business activities information

There are two categories of indicators to measure information on business activities of enterprises. The first is the business activity status reflected in the financial statements. The second is the business activity status reflected from the capital market. Financial characteristics are mainly composed of financial ratios and cash flow indicators. Performance in the capital market is mainly composed of stock price characteristics indicators.

The first type of indicators is mainly composed by financial ratios. For example, Altman (1968) selects 22 financial ratios from 5 aspects of liquidity, profitability, financial leverage, solvency, and operational efficiency for modeling and analysis. Finally, working capital/total assets, retained earnings/total assets, pre-tax profit/total assets, equity market value/debt book value and sales /total assets these five financial ratios are selected to establish the Z-score model.

Altman et al. (1977) build an index system with 7 financial ratios including return on assets, income stability, solvency, cumulative profitability, current ratio, capitalization rate and firm size, and construct the ZETA model which is suitable for large listed companies. Doumpos et al. (2015) collect data on 1,325 listed companies in Europe, covering 8 countries and 5 sectors, and construct a credit risk evaluation index system including profit/total assets, EBIT/interest expenses, total assets/ equity, equity/long-term debt, as well as other structured information such as company size, country risk (macro factors).

In addition to the above financial ratios, cash flow indicators are also considered to have an impact on a company's credit risk. Blum (1974) finds that the cash flow/total liabilities can effectively explain the credit risk of enterprises. Charitou et al. (2004) find that operating cash flow can effectively predict the failure of the company. Rujoub et al. (1995) conduct a five-year empirical study of 33 failed companies and 33 non-failed companies. In the final model, 8 cash flow ratios and 6 traditional financial ratios are used. It is found that the cash flow ratios have a higher predictive effect on the financial failure of the enterprise than traditional financial ratios. The combination of the two can significantly improve the prediction accuracy of the model. However, Casey and Bartczak (1985) show that the company's operating cash flow cannot significantly explain its credit risk. Gentry (1985) also obtains similar conclusions, but finds that changes in cash dividends can effectively predict corporate financial crisis.

The second category is capital market performance indicators of enterprises. Beaver (1968) first uses stock market yields to predict corporate credit risk. Aharony et al. (1980) incorporate the variance of market yields to the credit risk assessing index system and find that the stock market yields of bankrupt companies fluctuate significantly. Altman and Brenner (1981) and Clark and Weinstein (1983) also find that capital market performance can predict corporate bankruptcy.

The above index system is mainly aimed at large enterprises, especially listed ones. There are also literatures that use the above financial indicators when assessing the credit risk of SMEs. For example, Man et al. (2018) use Lasso-Logistic to study the influencing factors of credit risk of SMEs based on the data of 496 unlisted enterprises. They find that financial factors such as size, solvency and operational capabilities of SMEs have a significant impact on their credit risk. B. Zhou (2021) studies the Internet financing of SMEs. The credit evaluation index system constructed included hard information to measure the business activities of enterprises. It mainly covers four aspects: solvency, profitability, operating capacity and growth capability. Among them, solvency mainly measures whether an enterprise has the ability to repay short-term debt, long-term debt and interest. It mainly consists of current ratio (current assets/current

liabilities), quick ratio (quick assets/current liabilities), ratio of operating cash flow to liabilities, asset-liability ratio (total liabilities/total assets), equity multiplier (total assets/total shareholders' equity) and interest coverage ratio (EBIT/interest expense). Profitability is closely related to solvency. The stronger the profitability, the more guaranteed the future solvency. Profitability is mainly measured by net profit/total revenue, net profit/total assets and net profit/shareholders' equity. Operating capacity mainly measures the turnover efficiency of related assets, which is also closely related to solvency. The faster the turnover of various assets and the stronger the liquidity, the stronger the solvency of the enterprise. Operational capacity is mainly determined by inventory turnover (sales cost/inventory), accounts receivable turnover (sales revenue/accounts receivable), current asset turnover (sales revenue/current assets) and total asset turnover (sales revenue/ total assets). Growth capability mainly measures the development prospects of an enterprise. A company with better development prospects has a stronger ability to pay its debts in the future. Growth capability is mainly measured by two indicators: net profit growth rate and total asset growth rate.

However, some studies have also shown that due to the risk sources, financial information standardization and reliability of SMEs are different from those of large enterprises. Using purely financial indicators, the prediction of the credit risk of SMEs is low. For example, Edmister (1972) chooses financial ratio indicators, such as debt ratio, current ratio, return on equity and asset turnover ratio, to evaluate the credit risk of small enterprises, which reduces its prediction accuracy to a certain extent.

2.3.1.2 Financial behavior information

The financial behavior of firms and borrowers is also structured hard information. B. Zhou (2021) points out that, in the credit evaluation index system constructed for SMEs, the financial behavior information of enterprises includes accounting information and third-party rating information and credit records of SMEs. Among them, the accounting information mainly depicts the scale of various financial indicators such as the company's assets, liabilities, accounts receivable, and inventory, historical financing performance rates, settlement information, and POS transaction information. Credit records include default records and tax credit records. The third-party rating information mainly refers to the credit rating.

When Emekter et al. (2015) study the default risk of P2P, they also consider the financial behavior of borrower. Using data from online peer-to-peer (P2P) platform "Lending Club", they find that borrowers' credit ratings, debt-to-income ratios, FICO scores and utilization of revolving loan lines are all related to their probability of default. The lower the credit rating,

the higher the default rate. The longer the loan term, the higher the default rate. The higher the income and FICO score, the lower the default rate. Jiang et al. (2017) use the data from the P2P platform “Winglong Loan” and find that the borrower’s commercial insurance, credit card limit, number of failed bids, borrowing interest rate, repayment period and repayment method have a significant explanatory power for their default risk. The more commercial insurance purchased, the lower the default risk. The higher the credit card limit, the higher the default risk. The higher the number of failed bids, the lower the default risk. The higher the loan interest rate, the higher the default risk. The longer the repayment period, the lower the default risk. Ding et al. (2019) study the impact of borrowers’ historical performance on their default risk. The research finds that a borrower’s historical performance is a measure of their loan reputation, which is negatively correlated with the borrower’s default probability. The better the borrower’s reputation, the lower the default probability.

In theory, financing costs should be positively related to default risk. Jin et al. (2022) study the influencing factors of financing cost of SMEs under the premise of government credit enhancement. They conduct an empirical study using a sample of 4,502 loans from the SME Credit Guarantee Fund in Wenzhou, China, from 2017 to September 2020. The study finds that the loan amount, loan term, whether it is a high-tech enterprise, whether it is the first time to apply to a bank loan and whether it is a start-up enterprise all have significant relationships with the loan cost of SMEs. The larger the loan amount and the longer the term, the lower the financing cost of SMEs. The impact of the first loan on interest rates varies by loan issue time and by business. For loans issued before the second quarter of 2019, the financing cost of the first loan is lower, and while for loans issued after that, the financing cost of the first loan is higher. At the same time, for high-tech enterprises and start-ups, the financing cost of the first loan is relatively high.

Studies have shown that borrowing channels can significantly affect the default probability of borrowers. Tra Pham and Lensink (2008) use Vietnamese household survey data to study the effect of borrowing channel on the probability of default. They divide borrowing channels into three categories: the formal channel, semi-formal channel and informal channel, and find that borrowers from informal channel have a significantly higher default probability than borrowers from formal and semi-formal channels.

2.3.2 The application of soft information in credit evaluation

B. Zhou (2021) points out in the credit evaluation index system constructed for SMEs that the

soft information of enterprises is divided into four categories. It Includes the basic information, competitiveness information, social relations information and financing statement information. This thesis argues that moral hazard information is also a kind of soft information of enterprises, which has a significant predictive ability for the default behavior of SMEs (Castillo et al., 2021). The relevant literature on these five aspects is summarized below.

2.3.2.1 Basic information

The basic information includes the basic information of the enterprise, the information of enterprise's owner, and the operation warning information (B. Zhou, 2021). The basic information of the enterprise includes the number of employees, industry characteristics, main business type, geographical location and registered capital scale. The basic information of enterprise's owner includes the gender, age, marital status, income, education, political status and personal assets. The information on enterprise operation early warning mainly includes whether the enterprise has lawsuits, violations of regulations, and serious violations of the law or dishonesty. Scholars have studied and confirmed the predictive ability of basic corporate information on default risk from different perspectives.

For the information of operation early warning, Yin et al. (2020) study the predictive ability of legal judgments on the credit risk of SMEs. They find that deriving features from valid legal judgments significantly improves the model's classification ability compared to using only financial and corporate features. This thesis will summarize the existing literature from two aspects: the basic information of the enterprise and the basic information of the owner.

(1) Basic information on the enterprise

Fama et al. (2003) add the enterprise scale variable to the enterprise's credit risk evaluation model. They find the larger the scale of the company, the stronger the risk resistance, and the lower the loan default probability. Man et al. (2018) use Lasso-Logistic to study the influencing factors of credit risk based on the data of 496 unlisted SMEs. They find that the basic characteristics of enterprises such as their age (duration), the number of employees and other indicators have a significant impact on the credit risk of Chinese SMEs. Jia and Shi (2003) consider the scale and registered capital of enterprises in the credit risk factors of SMEs.

Cheng and Yang (2022) study the influencing factors of corporate loan rejection based on the survey data of small and micro enterprises in China in 2015. They find that the size, duration, and nature of the actual controller (or owner) of the firm have significant impacts on the likelihood of loan rejection. Specifically, larger enterprises and those with a state-owned background are less likely to have their loans rejected. Firm age has a non-linear relationship

with rejection probability. When the company has been in existence for less than 20 years, the older the company, the lower the probability of being rejected, but when the company has been in existence for more than 20 years, the older the company, the higher the probability of loan rejection.

(2) Basic information on the owner of the enterprise

Because many small and micro enterprises are sole proprietorships, partnerships, and individual industrial and commercial households, the operating conditions, financing capabilities, and repayment behavior of the enterprises are greatly affected by the actual controller or business owner. Online P2P lending is very similar to SMEs financing, both of which are difficult to provide banks with sufficient and accurate hard information to assess their default risk. Among the various soft information used by banks, the characteristics of business owners of SMEs and the characteristics of borrowers of P2P are important indicators. Among them, the race, gender, age, marital status, education level, income level, appearance, geographical characteristics, social network and loan reputation of business owners all have a significant impact on their financing probability and default probability. Much literature studies and confirms the impact of personal characteristics on firm and individual financing probability and default risk.

The first is the effect of race on default risk. Pope and Sydnor (2011) find that black people have a higher probability of default compared with white people. Therefore, in the P2P platform, the probability of successful loan of black people is 25%-35% lower than that of white people.

The second is the effect of appearance on financing probability and default risk. Ravina (2019) finds that the appearance of a borrower significantly affects the probability of loan success and the probability of default. The study finds that beautiful borrowers have a higher probability of loan success, but also a higher probability of default, relative to unbeautiful borrowers. Duarte et al. (2012) also study the effect of appearance on lending behavior. They find that borrowers who appeared to be more trustworthy are more likely to get a loan, while in fact they have higher credit scores and lower default rates.

The third is the influence of age and marital status. C. Wang et al. (2019) find that older people have a lower probability of default on P2P platforms, and married borrowers are more likely to obtain loans and have a lower default probability.

The fourth is the effect of gender. Barasinska and Schäfer (2014) conduct a study on P2P lending in Germany and find that female borrowers are more likely to obtain loans, which is inconsistent with the conclusions of literature research in traditional banking. D. Chen et al. (2017) conduct a study on China's P2P and find that gender discrimination also exists in China's

online P2P credit market. Consistent with Barasinska and Schäfer (2014), they find that female lenders have easier access to funds but pay higher borrowing rates. At the same time, they also find that the default probability of female borrowers is significantly lower than that of male borrowers.

The fifth is the influence of occupation, professional title and income level. Jiang et al. (2017) study the data based on the P2P platform “Winglongdai” and find that the borrower’s occupation and title have a significant impact on the probability of default. The higher the title of the borrower, the lower the default risk. Research by (C. Wang et al., 2019) finds that borrowers with higher incomes are more likely to obtain loans and have a lower probability of default.

The sixth is the effect of education level. Studies have shown that traditional financial institutions have “educational discrimination” when lending to individual investors. That is, they are more likely to lend to borrowers with better education and better school reputation. Some scholars have found that this kind of “educational discrimination” still exists in Internet finance. B. Yao (2016) studies whether the education level of borrowers affects their financing probability based on the loan data of China’s P2P platform “Renrendai”. He finds that lenders show different preferences for the education level of borrowers under different circumstances. When the credit evaluation mechanism is not perfect, they have obvious “educational discrimination” and prefer borrowers from more educated. When the credit evaluation mechanism is gradually improved, the influence of the borrower’s educational background on the probability of financing success becomes less obvious, and the “educational discrimination” gradually disappears. J. Li and Hu (2019) study the effect of borrower education on their P2P default risk. They collect relevant data from the platform “Renrendai” and find that the borrower’s university reputation was significantly negatively correlated with its P2P default risk, and significantly correlated with its loan success probability.

The seventh is the influence of personal psychological factors on default risk. Kai Yang et al. (2022) study the impact of personal social media information on credit risk assessment. They mine personal psychological factors from written texts posted on personal social media, and use data from P2P lending companies for analysis. They find that personal psychological factors helped to assess their credit risk.

The eighth is the influence of geographical characteristics on the financing probability and default risk of enterprises and individuals. Many studies have shown that financial institutions have “geographical discrimination” when reviewing loans. They prefer to lend to borrowers in more developed regions. Agrawal et al. (2011) argue that crowdfunding borrowers have a very

diverse geographic distribution. They use the data of the crowdfunding platform Sellaband from August 2006 to September 2009 to study the influence of the distance between musicians who initiate crowdfunding on the platform and entrepreneur investors on the probability of successful financing, and explore the role of geographical location in the crowdfunding in entrepreneurial projects. After controlling for the influence of entrepreneurs' offline social networks, they find that geographic characteristics are not a significant factor for the success rate of musicians' online financing, and the Internet eliminates the geographical barriers of financing.

Nonetheless, many studies on the Chinese market contradict their conclusions. For example, H. Lu et al. (2020) study the influence of geographical characteristics on default risk. They find that the higher the level of social capital in the region where the borrower is located, the higher the probability of successful loan application and the lower the probability of default. The social capital level of the borrower's area has the ability to predict the borrower's loan repayment behavior. C. Wang et al. (2019) also find that borrowers living in economically developed areas are more likely to obtain loans and have a lower probability of default. Jiang et al. (2017) use the data based on the P2P platform "Winglongdai" and find that the type of borrower's residence has a significant explanation for the default risk. B. Zhou (2021) believes that this is because, when compared with Western countries, China's Internet development history is relatively short, and the development of Internet finance has a relatively late start. At present, China has not yet formed a complete credit evaluation mechanism and credit reporting system, so there is more likely a regional preference in online P2P lending. As the research of Y. J. Wang and Sheng (2012) show, the geographic location of a firm has a significant impact on its financing success probability. usually. Businesses from developed regions have easier access to financing.

Claessens et al. (2008) discuss the sources of geographic and educational discrimination in financial markets. They believe that the root causes are the policy preferences of financial institutions or lenders, the information asymmetry between investors and financiers, and the imperfect credit evaluation mechanism of the market. In the case of unreliable and insufficient information on other characteristics, regional characteristics and educational characteristics are still important credit risk evaluation indicators for financial institutions and investors. The research of B. Yao (2016) supports the view of Claessens et al. (2008) to a certain extent. B. Yao (2016) collects 163,800 loan records of China's P2P platform "Renrendai" from January 2012 to May 2015 to study whether the geographical characteristics of borrowers affect their financing probability. He finds that lenders show different preferences for borrowers' geographic locations in different situations. When the credit evaluation mechanism is not

perfect, they have obvious “regional discrimination” and prefer borrowers from developed regions. When the credit evaluation mechanism is gradually improved, the influence of the borrower’s location on the probability of successful financing becomes less obvious, and the “regional discrimination” gradually disappears.

In addition to SME loans and P2P loans, financial institutions often rely on the personal characteristics of household heads when extending loans to farmers. The extraction and use of household head characteristics by financial institutions are basically consistent with the above-mentioned characteristics. It mainly includes the age of the head of the household, the duration of the family, the education level of the head of the household, family assets and income, the number and structure of the family population, ethnic categories, and occupational characteristics. At the same time, religious belief is also an influential factor that financial institutions focus on. For example, Ifeanyi and Blessing (2012) use 110 cooperative farmers in selected villages in Ogun State, Nigeria as the analysis objects, and study the influence of the respondents’ socioeconomic characteristics, demographic characteristics and other indicators on the repayment situation. Age of the head of the household is found to have a significant effect on repayment rates. Farmers age 20-40 had the highest average repayment rate at 69%. The higher the loan size and the higher the farm size, the higher the repayment ability. But the larger the family size, the lower its repayment ability. Arene (1992) uses data from the Supervised Agricultural Credit Scheme in Anambra State, Nigeria, to study factors influencing loan repayment rates for beneficiaries of smallholder maize farmers. It is found that farm size, household income, age of household head, agricultural experience, distance from home to the source of the loan, education level, household size, income level, and adoption of innovative technology all have a significant impact on the loan repayment rate.

In their research on the Chinese market, Q. Li (2018a) take farmers in poverty-stricken areas as the research object. Based on the survey data in Xinjiang in 2016, the three-stage simultaneous Probit model is used to study the influencing factors of farmers’ credit behavior in poverty-stricken areas. They find that the ethnicity of the household head, household income and property, and whether the loan is a co-guaranteed loan have a significant impact on the credit default of farmers. Specifically, when the household head is Han nationality, the default risk is lower. The higher the household income and the more property, the lower the default risk. When the loan is a co-guaranteed loan, the default risk is lower. Among the factors affecting the probability of financing, the proportion of household labor force, population size, income and finance have significant positive effects.

2.3.2.2 Competitiveness information

The information of enterprise competitiveness includes the industry competitive position, innovation investment and ability, and policy support given by governments at all levels (B. Zhou, 2021). Among them, industry competitiveness is mainly characterized by the market share of enterprises in the industry. The innovation investment and innovation capability are mainly described by indicators such as R&D, the number of patents, the value of intangible assets, customer evaluation, the proportion of new products and the order turnover rate. The policy support given by governments at all levels is mainly measured by the local GDP growth level, market environment and financial environment, whether the local area is a key policy support area, whether the industry in which the enterprise is located is a key policy support industry, and the financial support given by the government.

Based on the survey data, Scott (2004) discusses the impact of the company's market competition intensity, relationship with banks and other indicators on the company's default risk. A. N. Berger and Udell (2002) believe that the relationship between enterprises and banks is a kind of soft information held by banks. By mastering this soft information, it can not only help enterprises to provide financing success rate, but also help banks reduce loan risks. Lussier and Pfeifer (2001) use data from companies in Central and Eastern Europe to study the effect of competitiveness information the credit risk of SMEs. They comprehensively measure the competitiveness of enterprises with indicators such as the management level of the enterprise (such as staff allocation, education level, the use and planning of professional advice), the economic cycle in the macro environment, and the product life cycle in the industry environment. The study finds that factors such as enterprise management level and human resource factors not only have the ability to predict the credit default of American companies, but also have the ability to predict the financial default of SMEs in Central and Eastern Europe. Q. Li (2018b) believe that the intensity of competition differs across industries, and the factors affecting the default risk of different industries should also be different. Jia and Shi (2003) consider factors at the enterprise level when commercial banks evaluate the credit risk of SMEs. They mainly include the product quality, market stability, technical strength of SMEs, whether to change the purpose of the loan, management mode, whether it will transfer property, whether there is excessive investment, whether it can evade bank debts, whether the data is sufficient, whether the information is distorted, and whether the regulations system is perfect. These factors mainly reflect the competitiveness of enterprises.

At the same time, the quality of the business owner is also a reflection of the competitive

advantage of enterprises. Y. Zhao (2013) chooses three types of indicators for the quality of business owners, the type of business management and the dynamic risk control of banks to construct a credit risk evaluation system for small and micro enterprises. Among them, the quality of business owners includes three aspects: personal quality, social reputation and management ability.

2.3.2.3 Social relationship information

The social relationship information of an enterprise includes two aspects: the upstream and downstream relationships of the enterprise and the social network relationship of the enterprise (B. Zhou, 2021). The upstream-downstream relationship refers to the company's stock holdings and guarantees. The social network relationship mainly refers to the performance of the company on the e-commerce platform, such as whether to open an online store, the credit rating of the online store, and its performance on public media, such as whether the company has opened a social media account, and the number of articles published on social platforms, the number of followers on social platforms. The mechanism by which the social network relationship of an enterprise affects its financing credit lies in two aspects. On the one hand, social networks can bring benefits to enterprises, such as increasing the popularity through public attention, increasing product sales, and thus enhancing their repayment ability. On the other hand, the social network will form a supervisory role for the behavior of the enterprise, and form constraints on the enterprise through the social reputation and word of mouth of the enterprise, thereby increasing its financing credit.

(1) The impact of social capital on loan credit

Similar to the social relations of enterprises, the owner of the enterprise and individuals involved in P2P lending also have social relations, and their social relations will also affect their financing success probability and default probability. Karlan (2007) finds that borrowers can effectively assess their credit risk through the social relationships of the poor. Using data from FINCA-Peru, using residential distance and cultural similarity to describe the strength of social ties, the study finds that poorer people with stronger social ties have higher repayments and savings rates through randomized experiments. The reason is that social connections can form a very strong supervisory role for individuals, making them afraid to breach the contract. M. Lin et al. (2013) study the impact of borrowers' online friendship on their financing availability and default probability on P2P platforms. The study finds that borrowers with online friendships are more likely to obtain financing, have lower interest rates on their loans, and have a lower risk of default. This indicates that the quality of a borrower's social network is predictive of the

default risk.

“Group” is a kind of social network on P2P platform. Group relationships are also one type of social capital owned by enterprises and individuals. For example, the P2P platform Prosper has established a complete group system. The system allows and encourages borrowers and lenders to form group relationships, with groups acting as an intermediary between the two. The “group” mainly affects the borrower’s credit through the following channels. Groups will have a binding effect on borrowers and improve their performance motivation (Granovetter, 1985); groups with a good reputation have more financing opportunities, and group members will influence and supervise each other, and encourage group members to perform (Ba & Pavlou, 2002); borrowers who join the group generally have long-term financing plans (Greiner & Wang, 2010); the group leader has a strong supervisory role and high credit guarantee, and the group acts as a financial intermediary to effectively reduce information asymmetry between borrowers and lenders can significantly improve borrowers’ credit status (S. C. Berger & Gleisner, 2009).

Some literature uses the transaction data to study the effect of “group” on the success probability and default probability of individual financing. For example, Peng and Yang (2017) base on the transaction data of the P2P platform Prosper, study the influence of banks, enterprises, and individuals’ grouping behaviors and alliance behaviors on borrowers’ financing probability, borrowing cost, and default probability. Their empirical study finds that borrowers who tend to form alliances generally have lower credit ratings, and that they significantly increase the probability of financing and reduce borrowing rates by joining alliances. But borrowers who join groups and alliances have a higher risk of default than borrowers who do not. Pei and Caoi (2016) also use 126,000 transaction data from the P2P platform Prosper to study the impact of joining a group on the borrowing success rate and default risk. Greiner and Wang (2010) regard “group” as a trust mechanism, establish a theoretical model and conduct empirical research using the transaction data of P2P platform Prosper, and discuss the role of trust mechanism in credit evaluation in P2P credit. Freedman and Jin (2008) use transaction data from the Prosper.com platform from June 1, 2006 to July 31, 2008 to study the impact of social networks on borrowers’ financing behavior and credit risk. They find that cohort loans had lower yields than non-cohort loans because of their lower risk of default, but this yield gap is gradually narrowing. The reason is that Prosper removes the incentives for group leaders, making it possible for them to fund low-quality loans. Using data from six different loan programs in Madagascar, Zeller (1998) studies the impact of group loan program design and group characteristics on loan defaults. The research finds that groups can diversify their asset

portfolios and risks, reduce default probability, and improve repayment performance. In a study of the Chinese market, Qinlan and Izumida (2013) use a Logit model to explore how borrower- and group-level characteristics affect the repayment decisions of group borrowers, based on the survey data 245 rural collected from Rural Credit Cooperatives (RCCs) in Guizhou Province Southwest China. The study finds that factors such as the level of acquaintance in the group and whether or not they are employed in a government agency significantly increase borrowers' repayment rates.

The bank's trust in the enterprise is also a kind of social relationship that the enterprise has. Using data from 160 SMEs in Finland, Kautonen et al. (2020) examine the impact of trust on corporate credit availability. The study finds that there is a significant positive correlation between trust and firms' access to credit. The higher the bank's trust in the enterprise, the higher the financing probability of the enterprise. It shows that trust as a kind of informal soft information has substitutability with other formal information of the enterprise.

The above studies all believe that social relations can effectively reduce the default rate of borrowers and improve their repayment ability and repayment willingness. However, some studies have found that social relations may have a negative impact on borrowers' performance behavior. For example, Freedman and Jin (2017) study the influence of borrower's social relationship on their default risk. They find that borrowers with social ties are more likely to default, and social ties do not measure borrowers' credit quality, contrary to the findings of M. Lin et al. (2013). Ahlin and Townsend (2007) study the impact of joint liability loans on repayment rates using data from Thai borrowing groups. They find that when the joint liability ratio is higher, the repayment rate is lower, and when there are more social connections, the repayment rate is lower.

Some literature has found that different types of social relations have different effects on financing behavior. Q. Li (2018b) use the Probit model to empirically study the impact of formal and informal social capital on the credit default risk of farmers, based on the farmer household survey data in Jiangsu and Shandong provinces of China in 2013. The number of Chinese Communist Party members in a family is used to measure the formal social capital of the family and to characterize the social relations brought about by institutionalized factors. The key payment in the loan year is used to measure the non-social capital of farmers, and to describe the social relationship brought about by non-institutionalized factors. Their research find that the amount of formal social capital is significantly negatively correlated with credit default risk. The more the formal social capital, the lower the default risk of farmers. The amount of informal social capital is significantly positively correlated with the credit default risk of farmers. The

more informal social capital, the higher the default risk of farmers.

Some literature has found that different levels of social relations have different effects on financing behavior. Petersen and Rajan (1994) empirically study the influence of the relationship between the company and the bank on the financing probability and financing cost. They find that when companies have a close relationship with one bank, it will significantly increase their probability of financing success, but have a smaller impact on financing costs. But at the same time, if the company has established close relationships with multiple banks, it will have a negative effect on their financing success probability and financing cost.

(2) The influence mechanism of social capital on loan credit

Part of the literature discusses the mechanism by which social relations affect corporate and individual financing credit, mainly in two aspects. One is the positive promotion. Social network will improve the repayment ability and repayment willingness of enterprises and individuals. The second is the reverse promotion. The social network will restrict the repayment behavior of enterprises and individuals through social image and reputation, and reversely encourage them to perform their contracts.

In terms of theoretical research, A. X. Li (2014) studies the impact mechanism of social capital on the credit risk of farmers. This study proposes that social capital can be divided into two categories: cognitive social capital and relational social capital. Among them, cognitive social capital mainly has institutional attributes, while relational social capital mainly has capital attributes. Institutional attributes mainly play a restraining function, while relational attributes mainly play an incentive function. Institutional attributes will affect farmers' repayment willingness through supervision mechanism, reputation mechanism and punishment mechanism, while capital attributes can improve farmers' repayment willingness by obtaining credit resources and income. At the same time, social capital will also increase farmers' willingness to repay loans by increasing their income, broadening their credit channels, and improving farmers' ability to deal with environmental risks.

In terms of empirical research, the research of Q. Li (2018b) find that the formal social capital held by farmers will improve the repayment ability and repayment willingness of farmers through the supervision mechanism, thereby reducing the default risk of farmers. Although informal social capital helps to enhance the repayment ability, it significantly reduces the repayment willingness of farmers, thereby increasing the default risk of farmers.

2.3.2.4 Financing statement information

In the traditional offline financial market, banks and other financial institutions will pay close

attention to the borrowing purposes of enterprises and individuals when reviewing loans. The rationality of use has a significant impact on the success rate of financing. In the online Internet financial market, companies and individuals will also choose whether to provide loan statements voluntarily when financing. Although borrower representations are cost-free, voluntary, and unverifiable unreliable information, they can reduce borrowing costs by 1.27% while increasing the probability of financing success by 7% (Michels, 2012). The information of the financing application statement includes the characteristics of statement text and subject, the expressed repayment ability and repayment willingness, and the expression of financing emotions (B. Zhou, 2021). In this thesis we will review the impact of different characteristics of financing presentation on default risk and financing probability.

(1) The subject, length, misspellings and repetition of the financing statement

Jiang et al. (2017) randomly select 15,806 loan records from January 2014 to January 2016 on the “Winglongdai” platform, to study the influence of the distribution of subject words in borrowers’ statements on their default probability. The total sample contains 9006 non-default records and 6800 default ones. They separately extracted six subject headings from borrowing statements: assets, income, work, family, business, and agriculture, and one feature: text length. Using logistic model regression, they found that the frequency of asset themes in statements is significantly negatively correlated with default risk, that is, the more assets described and the higher the frequency, the lower the default risk of the borrower. The frequency of other topics such as income, work, and family in the statement is significantly positively related to default risk. Statement text length is significantly negatively correlated with default risk, therefore, the longer the statement, the lower the default risk.

L. Chen et al. (2019) conduct a similar study. They collect borrowers’ statements from P2P platform “Renrendai”, and manually extract information such as text features and emotional features. Empirical research finds that the length of loan statements is negatively correlated with default risk, and the number of repeated statements is positively correlated with default risk. The more fully the repayment willingness expressed in the loan statement, the smaller the default risk of the borrower. The more urgent the loan demand, the higher the default risk.

Larrimore et al. (2011) study the influence of language features in loan statements on the probability of loan success. They find that the more statements related to one’s financial situation, the easier for him/her to gain the lender’s trust to obtain a loan. Ye and Chen (2017) also use the order data of the Chinese P2P platform “Renrendai” to study the effect of the number of words in the loan description on the borrower’s financing credit. They find that the word count of the borrowing statement is significantly positively correlated with the probability

of financing success, but significantly increases financing costs.

Dorfleitner et al. (2016) use data from two major European P2P platforms, Auxmoney (www.auxmoney.com) and Smava (www.smava.de), to study the impact of borrower representations on the probability of successful financing and the probability of default. They find that misspellings in the text, text length, and keywords reflecting positive sentiment, could predict the probability of financing but not the probability of default.

(2) The description of identity and personality in financing statement

Herzenstein et al. (2011) use data from the P2P lending website Prosper.com to study the impact of identity claims in borrower statements on the probability of financing success and repayment behavior. They find that the more identity statements in the borrower's statement, the higher the probability of successful financing, but also the higher the probability of default. The content of the borrower's identity statement also has a significant impact. Specifically, borrowers who present themselves as trustworthy or successful identities are able to obtain higher financing amounts, but have less predictive power of loan performance, relative to truthfully stated identity characteristics such as moral and financial difficulties.

H. J. Wang and He (2015) randomly select 2000 observations from orders with financing statements from March 1, 2012 to March 1, 2013 on the Chinese P2P platform "Renrendai", to study the influence of borrower personality on the financing success probability. They divide borrowers' personalities into six categories, namely "integrity", "success", "diligence", "steadfastness", "economic difficulties" and "affection". The study finds that when the personality in the borrowing statement is "diligent", it has the highest impact on the probability of financing success. The rest of the personalities are ranked according to the size of the marginal impact, namely "affection", "integrity", "steadfastness", "success" and "economic difficulties". "Success", "economic difficulties," and "diligence" personalities are significantly negatively associated with borrower default rates. Among them, the "Success" personality has the strongest ability to predict default behavior, followed by "economic difficulties" and "diligence". The remaining three personalities have no predictive power for borrower defaults.

(3) Punctuation marks in financing statements

X. Chen et al. (2018) use data from the Chinese P2P platform "Renrendai" to study the impact of punctuation use in borrower statements on the probability of financing and the probability of default. They find that more punctuation in a given word count decreases the probability of financing and increases the probability of default. Excessive use of punctuation can make loan descriptions informal and undermine lender trust. Ye and Chen (2017) also use the order data of the Chinese P2P platform "Renrendai" to study the impact of punctuation in

loan descriptions on borrowers' financing credit. They find that the number of punctuation marks is significantly negatively correlated with the probability of financing success, consistent with X. Chen et al. (2018). At the same time, it is also found that the number of punctuations is significantly positively correlated with the financing cost, and the more the number of punctuations, the lower the borrowing cost.

(4) Description of the loan purpose in the financing statement

Zhuang et al. (2015) expound the impact of investment bias in P2P borrowers' financing statements on financing probability and interest rate by constructing a theoretical model. They divided the loan purposes into four categories: investment, consumption, experience and others, and identified the borrowers as students, wage earners, online shop sellers, private owners and others. They select 196,720 loan orders from the P2P platform "Paipaidai" from 2014 to 2018 as a sample for empirical analysis. The regression results show that loan applications for investment purposes are more likely to be financed than for consumption purposes. At the same time, when borrowers are private owners, their investment loan needs are more easily supported, and when borrowers are students, their consumer loan needs are more easily supported.

2.3.2.5 Moral hazard information

Many scholars try to measure the magnitude of corporate moral hazard and study the impact of moral hazard on default risk. For example, José A Castillo et al. (2018) propose four characteristics of moral hazard. One is asset substitution, measured by $(\text{short-term investment} + \text{long-term investment}) / \text{total assets}$, the higher the indicator, the higher the moral hazard. The second is low effort level, measured by $\text{operating costs and expenses} / \text{total assets}$, the higher the indicator, the lower the moral hazard. The third is insufficient investment, measured by $\text{non-operating expenses} / \text{total assets}$. The higher the indicator, the higher the moral hazard. Fourth, the different uses of invested capital, measured by $\text{capital assets} / \text{total assets}$, the lower the index, the higher the moral hazard of the enterprise. Although they provide new indicators for SME credit assessment from the perspective of moral hazard, they do not establish a unified moral hazard index.

On their basis, José A. Castillo et al. (2021) construct a unified moral hazard index based on data from the Colombian database from 2007 to 2014, and examine the impact of the moral hazard index on the probability of corporate default. The study finds that adding moral hazard indicators on the basis of traditional financial indicators can significantly improve the model's ability to predict default behavior.

To sum up, due to the different credit risk characteristics of large enterprises and SMEs,

different evaluation index systems should be adopted. The evaluation of large enterprises and listed companies mainly relies on the open financial information and capital market performance information disclosed by the enterprises, which are basically quantitative indicators. While the evaluation of small and medium-sized enterprises, due to the lower reliability and completeness of their financial statements, depends on non-financial indicators to a certain extent, which are mainly reflected as qualitative indicators.

Table 2.1 summarizes the index system used in the evaluation of SMEs in the existing literature. As can be seen from the table, existing research also shows that the current evaluation of small and micro enterprises mainly relies on the information provided by the enterprise, and less on the information shared by third parties such as government. The existing literature is used to evaluate the credit risk indicators of SMEs, whether it is hard information or soft information, notice that almost no tax information of enterprises is involved.

Table 2.1 Summary of the credit risk evaluation index system for SMEs

Information category	First-level indicator	Second-level indicator	Third-level indicator
Hard information	Operating activity information	Solvency	Current assets/current liabilities
			Liquid assets/current liabilities
			Cash and equivalents/current liabilities
			Total liabilities/total assets
			EBIT/interest expense
		Operating capacity	Total assets/shareholders' equity
			Cost of sales/inventory
			Sales revenue/accounts receivable
			Sales revenue/current assets
			Sales revenue/total assets
		Growth ability	Net profit growth rate
			Total asset growth rate
			Net profit/total revenue
		Profitability	Net profit/total assets
			Net profit/shareholders' equity
			EBIT/total assets
			EBIT/total assets
		Cash flow ratio	retained earnings/total assets
			Operating Cash Flow/Total Liabilities
			Cash flow adequacy ratio
	Financing behavior information	Capital market performance	Cash flow reinvestment ratio
			Stock return
		Account information	Volatility in stock returns
			The total amount information in the financial statements of the enterprise, such as the total amount of assets, liabilities, inventory and other subjects
			Settlement information of the business in the bank account
			POS transaction information
			Whether to buy commercial insurance

Soft information	Basic information	Basic information of the enterprises	Credit record	Historical performance information Whether the company has debt default records Whether the company has tax default records
			Third party rating	Debt credit rating Tax credit rating
			Loan feature	Loan size loan term Lending rates loan channel Utilization of Revolving Loan Facility Repayment period Repayment method Whether it is the first loan
				Enterprise size Industry characteristics Geographical features Geographical features Main business type Age Number of registered capital Whether the enterprise have a state-owned background Whether the enterprise is a start-up Whether it is a high-tech enterprise
			Basic information of the business owner	Race Age Gender Marital status Educational level Nationality Religion Political status Appearance Occupation and title Income Credit card limit Geographical features Psychological factors
			Warning information	Whether the company has a lawsuit Whether the company has a lawsuit Whether the company has serious dishonesty
			Industry competitiveness management level	Company's market share The intensity of competition in the industry The level of human resources of the enterprise Enterprise's strategic planning ability
			Innovation level	R&D investment Number of patents Enterprise's intangible asset value Customer reviews Ratio of new products Ratio of new products
			external environment	macroeconomic cycle Product Lifecycle The economic environment, financial environment and other market environment

Social relationship information	Social relationship of the enterprises	Whether the location of the enterprise is a key policy support area
		Whether the industry is a key policy support industry
		Business investment
		Company's guarantee
		Whether the company has an online store
		Credit rating of online store
		Does the company open a social platform account
		Number of articles published on social platforms
		Number of followers on social platforms
		Whether the business is part of a financing group or consortium
Financing statement information	Social relationship of the business owner	Bank trust in business
		The relationship between business and bank
		Whether the business owner is a member of the Communist Party of China
		Social Networking and Social Relationships for Business Owners
		Statement length, word count
		Number of punctuation marks
		Number of repeated statements
		Text subject
		Borrower identity and personality description
		Description of loan purpose
Moral risk information	Moral hazard indicator	Description of willingness to repay
		Asset replacement
		Low effort level
		Low effort level
		Different uses of invested capital
		The (positive) effect of the moral hazard variable on the pd ratio
	Moral hazard index	

Source: Author.

2.4 Research on information sharing

In the credit market, due to the poor reliability of financial data provided by SMEs and individual borrowers, there is a problem of information asymmetry between borrowers and lenders. In order to accurately assess the credit risk of borrowers and avoid and reduce adverse selection and moral hazard caused by information asymmetry, banks generally need to dispatch loan officers to collect data through field investigations and obtain data from third parties.

2.4.1 Theoretical research on the impact of information sharing

This thesis will summarize the existing literature from three aspects: the impact of credit information sharing, the motivation for sharing credit information and the content of credit information sharing.

(1) The impact of credit information sharing

To the best of our knowledge, Pagano and Jappelli (1993) are the first to formally propose credit information sharing. They build an adverse selection model, taking information sharing among lenders as an endogenous factor, and study the factors of information sharing. They find that information sharing can effectively improve the liquidity and heterogeneity of borrowers, can significantly increase the size of the credit market, and promote the progress of information technology. Specifically, Pagano and Jappelli (1993) believe that information sharing can reduce the information asymmetry between borrowers and lenders, thereby reducing adverse selection, which will increase the number of people who get loans and the scale of credit. Meanwhile, the expansion of credit scale will reduce the operating cost of the credit reporting system, expand the number of credit-reported persons, promote information sharing in the reverse direction, further promote the expansion of credit scale, and generate network effects. Pagano and Jappelli (1993) also find that with the intensification of market competition, the benefits of information sharing decrease marginally.

Klein (1992) constructs a theoretical model based on game theory, and studies the influence of credit bureau's reputation on borrower's performance behavior. The study points out that the information sharing mechanism has economies of scale. When more creditors participate in information sharing, it will significantly increase the borrower's default cost and reduce the possibility of default.

Bennardo et al. (2015) find that information sharing can effectively avoid over-borrowing by longs in credit transactions. When there is no information sharing in the credit market, multiple banks may lend to the same borrower. This has led to over-borrowing by longs, making borrowers take on too much debt and exacerbating risks for banks. To avoid over-borrowing by longs, banks have adopted a system of credit rationing that limits the amount of credit given to a single borrower. When there is information sharing between banks, it can effectively avoid over-borrowing by longs, thereby reducing the credit rationing system and increasing the overall scale of credit.

(2) The motivation for sharing credit information

Padilla and Pagano (1997) discuss whether there are incentives to generate endogenous information sharing mechanisms among banks. When banks monopolize borrower information, borrowers have an incentive to reduce their effort levels to avoid high interest rates caused by information asymmetry. But if the bank shares borrower information with other institutions, will it have a positive impact on the bank's profits? Padilla and Pagano (1997) establish a two-period theoretical model and find that as long as banks retain an information lead in the initial

stage, their information sharing will increase borrowers' efforts, thereby increasing bank profits. There are two main paths through which information sharing affects banking theory. On the one hand, information sharing will effectively increase financial institutions' constraints on borrowers, reduce their moral hazard, and avoid credit defaults. This effectively increases the bank's profits. On the other hand, information sharing will reduce the cost of information acquisition between banks, increase competition within the industry, and reduce borrowing rates, which will reduce banks' profit margins. R. Zhang and Liang (2006) establish a credit information sharing decision model to analyze the motivation of Chinese banks to share credit information. The study finds that due to differences in the expectations of different banks, information sharing mandated by government departments may not only motivate banks to participate, but also may lead to bank dissociative behaviors.

(3) Content of credit information sharing

Padilla and Pagano (2000) study the differences in the binding force of borrowers on the content of information shared by creditors. They find that when too detailed information is shared among creditors, it reduces borrowers' incentives to perform. Compared with sharing all information, the creditor only shares the borrower's historical default information, which has a greater incentive effect on the borrower's performance. The study shows that creditors can improve borrowers' motivation to perform by adjusting the content of shared information.

2.4.2 Empirical research on the impact of information sharing

M. Brown et al. (2009) use corporate data to study the impact of information sharing among banks on the performance of credit markets in Eastern Europe and the former Soviet Union. Their research finds that information sharing can improve the probability of successful financing of enterprises, while reducing the financing cost of enterprises. The impact of information sharing varies from company to company and from country to country. For companies with less information transparency, and for countries with weak legal environments, the impact of information sharing on financing probability and financing costs is stronger.

Many literatures also discuss the difference in the impact of credit information from private and public information bureaus on credit markets. For example, Djankov et al. (2007) conduct research using private and public information data of 129 countries from 1999 to 2003, and find that the development of creditor rights protection legislation and information sharing institutions will increase creditor credit, which in turn leads to an increase in debt scale. Galindo and Miller (2001) collect data from public credit registries in 34 countries and enterprise-level

information from private credit registries in more than 30 countries. The study finds that both public information registration and private information registration can effectively reduce the financing constraint of enterprises. Through theoretical analysis, Jappelli and Pagano (2002) point out that information sharing among lenders will effectively increase the scale of loans and reduce the risk of default. Meanwhile, they conduct an empirical study using data from 43 non-European countries. Private credit bureaus and public credit bureaus are found to have complementary roles. Of the countries with public credit registries, only 30% have private information bureaus. In countries without public credit registries, 65% have private credit bureaus. They also find that both private credit bureaus and public credit bureaus' information sharing have the effect of reducing adverse selection and moral hazard.

2.4.3 Research on China's credit information sharing

Developed countries generally have a complete social credit system with rich credit data resources. Third parties can be private credit bureaus or government agencies. However, China's social credit system has just been established for a short period, and there is few public credit data available. Third-party data is mainly collected by the government, including corporate tax information, business information, and social security fee payment information.

J. M. Wang (2004) finds that China's basic credit information is mainly mastered by various government departments, and the degree of sharing is low. No matter if it is information sharing within government departments, or information sharing between government and social departments such as banks and enterprises, the level is relatively low. The dispersion of credit information has hindered the development of China's credit reporting industry, and has also led to the slow construction of China's credit system. Guo et al. (2013) study the status quo of credit information sharing among Chinese government departments through case studies, and analyze the existing problems. The study finds that although the Chinese government's credit information sharing has made some progress, there are still many problems. For example, communication barriers lead various departments to maintain their own rights and interests, and the sharing awareness is poor. Low sharing investment leads to imperfect sharing platform functions and low service levels. In order to solve the above problems, the study proposes that government should increase the awareness of sharing, establish a credit information database, improve the functions of the information sharing platform, and continuously develop new credit products.

H. Y. Zhang and Jiang (2008) establish a basic probability model to discuss the influencing

factors of credit information sharing among Chinese commercial banks. They analyze the liquidity of borrowers, differences in bank scale, and information sharing costs, and find that the probability of establishing a credit information sharing mechanism among Chinese commercial banks is relatively small, and it is more feasible for government departments to establish a public credit information system.

2.5 Literature summary

First, in terms of credit risk assessment methods, the industry and academia have developed a large number of methods and models for assessing corporate credit risk. With the development of artificial intelligence technology, more and more new data mining technologies are applied to credit risk assessment practices. However, not all models are suitable for micro, small and medium-sized enterprises. First, the multivariate discrimination credit risk assessment method represented by the Z-score and ZETA model mainly relies on financial statement information of the enterprises, but the financial information of micro, small and medium-sized enterprises has serious problems such as data missing, low standardization and unreliability. Therefore, these kinds of models are not suitable for assessing the credit risk of micro, small and medium enterprises. Second, models such as KMV and Credit Metrics rely on the market value of corporate assets and liabilities, which are mainly applicable to listed companies, and not suitable for non-listed companies, especially micro and small enterprises. Third, Random Forest is an integrated algorithm that requires a large enough sample size. In contrast to that, the banks' credit business volume of micro and small enterprise is relatively small, and the information stored is relatively incomplete, so the existing literature rarely uses Random Forest to evaluate the credit status of micro, small and medium-sized enterprises. Therefore, in the above credit risk evaluation methods, there are three main types of credit evaluation methods that are more commonly used in micro, small and medium-sized enterprises: Logit / Probit model, ANN and SVM.

Second, in the aspect of the credit risk evaluation index system, the existing literatures have studied the indexes affecting the enterprise credit risk from the two aspects of hard information and soft information. Among them, hard information includes corporate financial information and financial behavior information. While soft information includes basic information of the enterprise and the business owner, information on the competitiveness of the enterprise, information on the social relationship between the enterprise and the business owner, information on corporate and personal financing statements, and moral hazard information.

Each type of information contains multiple secondary indicators. For large and medium-sized enterprises, especially listed enterprises, existing literatures mainly rely on financial information and capital market information to assess their credit risk. For SMEs and individuals, their hard information and soft information will be comprehensively used, with special emphasis on soft information. However, we can find that most of the information, nonmatter hard or soft information, is provided by the enterprises themselves or obtained from financial research such as banks, and is rarely provided by government departments. The tax information of government departments is a very reliable hard information, and it is highly related to the real financial status of enterprises, but the current index system lacks the use of tax information.

Thirdly, in terms of credit information sharing research, existing literatures discuss the impact of credit information sharing, and empirically test the impact of information sharing by public information bureaus and private information bureaus on corporate and personal financing. In Western countries, both public credit investigation and personal credit investigation are relatively developed. In China, almost all credit information is in the hands of government, but the current level of government sharing of credit information is not high.

In practice, the Chinese government's tax department has shared the tax information of SMEs with banks, which can provide support for banks' credit risk assessment. However, banks have not made full use of this information, and existing studies have not used tax information as an important indicator for predicting corporate default risk. This thesis will add a series of tax payment characteristics of enterprises to the existing index system, hoping to improve the bank's loan review mechanism, and improve the approval rate and reduce the default rate by optimizing the credit risk assessment model.

Chapter 3: Research Methods

The main purpose of this thesis is to build a better loan review mechanism for “Tax E Financing”, so that companies applying for loans can be more accurately classified into “will default” and “will not default”. The essence is to find a better credit risk classification model, which contains two questions. The first is to select a feature set for classification, that is, to find a better credit evaluation index system for “Tax E Financing”. The second is to find a better classification algorithm. This thesis will use a variety of machine learning algorithms to optimize the credit risk evaluation index system, train and test the classification model based on index data, and select the optimal model as the credit risk classifier of “Tax E Financing”. The following will introduce the research methods from three aspects: model evaluation methods, feature processing methods and classification algorithms.

3.1 Model evaluation methods

3.1.1 Empirical error and generalization error

The difference between the predicted output of the classifier and the true output of the sample is called error. The error on the training set is named training error or empirical error, and that on the new sample is named generalization error. There are many classifiers that have very small training errors but large generalization errors, which is called overfitting, and the opposite is called underfitting. When choosing a classifier in practice, we hope to get a smaller generalization error. Therefore, a testing set is needed to test the classifier’s ability to discriminate new samples, and then the testing error is used as an approximation of the generalization error.

This thesis will comprehensively use the hold-out method, cross validation method and bootstrapping method to generate the training set and testing set from the original data. The hold-out method refers to the direct and random division of the data set into two mutually exclusive sets, one is used as the training set and the other as the testing set. The performance of the classifier on the testing set is used to evaluate its predictive ability. Cross validation method refers to dividing the original data set into k mutually exclusive subsets of similar size, each of which maintains the consistency of the data distribution as much as possible. Then each

time the union of $k-1$ subsets is used as the training set, and the remaining subset is used as the testing set, so that k training and testing can be performed, and the average of the k test results is returned as the performance measure of the model. This is also called k -fold cross validation. The bootstrapping method is based on the bootstrap sampling method (Tibshirani & Efron, 1993), which extracts m samples from the original data set D through sampling with replacement and puts them in the data set D' as the training set, and then use $D \setminus D'$ as the testing set.

3.1.2 Performance measurement

(1) Precision, Recall and F1

Performance measurement refers to the criteria for evaluating the generalization capabilities of classifying models. When comparing the capabilities of different models, using different performance metrics will often lead to different results. The following will introduce the performance metrics used in this thesis.

The problem studied in this thesis is essentially a two-class problem, that is, how to divide the companies applying for loans into two types that “will default” (labeled as +1, which is the positive sample mentioned below) or “will not default” (labeled as -1, which is the negative sample mentioned below). Therefore, according to combination of the true labels and predicted labels of the classifier, the samples can be divided into four categories that true positive (TP), false positive (FP), true negative (TN), and false negative (FN). $TP+FP+TN+FN$ = total number of samples. Expressing the above combination in the form of a table will result in a confusion matrix, as shown in Table A2.

From the confusion matrix, we can get four indicators: Accuracy, Error rate, Precision and Recall. The relevant formula is as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.1)$$

$$Error_rate = \frac{FN+FP}{TP+FP+TN+FN} \quad (3.2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3.3)$$

$$Recall = \frac{TP}{TP+FN} \quad (3.4)$$

Among them, Accuracy refers to the proportion of samples that are correctly classified, also known as correct rate or accuracy rate, and Error Rate refers to the proportion of samples that

are classified incorrectly. These two indicators are used to evaluate model performance in most classification tasks. However, they cannot meet all the task requirements of the problem studied in this thesis, because the model studied in this thesis is used to evaluate the company's possibility of default. But in fact, defaulting companies are only a small part of the companies applying for loans. For example, among the 600 non-defaulting and defaulting samples collected in this thesis, there are only 56 defaulting samples, accounting for less than 10%. At this time, even if all companies are classified as "will not default", the accuracy of the model can reach more than 90%, but all the defaulting companies have not been identified.

For banks, there are two goals in loan review. The first is to identify companies with high default risk as much as possible, and reject their loan applications to reduce the bank's bad debt rate. The second is to identify companies that will not default as much as possible, and grant them loans, which can increase the interest income of banks, enhance the financing success rate of SMEs, and promote the country's economic development. At this point, we need to use Precision and Recall as the performance measurement of the credit risk classification model.

Precision refers to the proportion of the samples classified as positive by the model that are truly positive, that is, how many companies that the model predicts "will default" actually default. Recall refers to the proportion of the truly positive samples that are classified as positive by the model, that is, how many defaulting companies are predicted by the model as "will default". Precision and Recall are a pair of contradictory indicators. To increase Recall, it is necessary to judge more companies as "will default", but this will increase TP and FP, resulting in a decrease in Precision. Therefore, there is a trade-off between Precision and Recall. F1 is defined based on the harmonic average of Precision and Recall, and its formula is shown below.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3.5)$$

(2) ROC curve and AUC

Many classifiers will generate a real value or a probability prediction for the testing sample, and then compare this prediction value with a threshold. If the predicted value of a sample is greater than the threshold, it is classified as positive, otherwise negative. Different thresholds will produce different classification results and performances.

Receiver Operating Characteristic (ROC) originated from the radar signal analysis technology used for enemy aircraft detection in the "World War II". It has been used in psychology and medical testing applications in the 1960s and 1970s, and has since been introduced into the field of machine learning (Spackman, 1989). The ROC curve uses the true positive rate (TPR) under different thresholds, which is Recall defined in the previous section,

as the vertical axis, representing the ratio of true positive samples classified as positive by the model. It uses the false positive rate (FPR) as the horizontal axis, which represents the ratio of the truly negative samples incorrectly classified as positive. The formulas are as shown in (3.6) and (3.7).

$$TPR = \frac{TP}{TP+FN} \quad (3.6)$$

$$FPR = \frac{FP}{TN+FP} \quad (3.7)$$

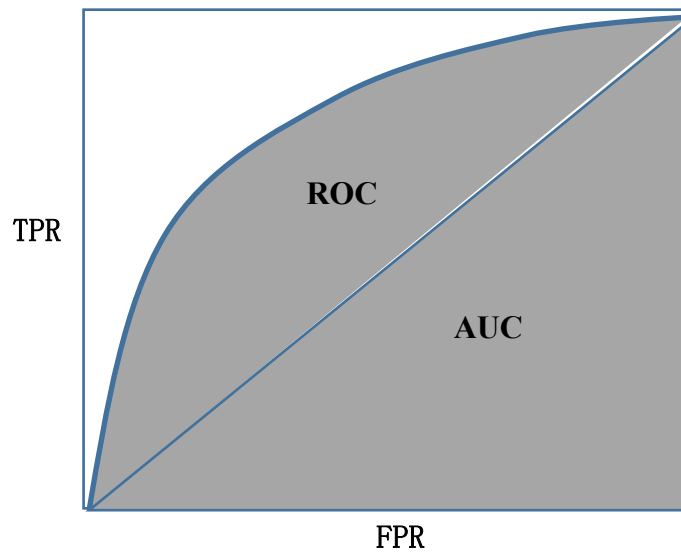


Figure 3.1 ROC curve and AUC

Source: Author.

The drawing process of ROC curve is as follows. Given m^+ positive samples and m^- negative samples, we sort the samples according to the prediction values (probability or real value) of the classification model. Firstly, the maximum value is set as the threshold, and all samples are predicted as negative. At this time, TPR and FPR are both 0, and the ROC curve is at the coordinate (0, 0). Then, the classification threshold is sequentially set as the predicted value of each sample, and the new TPR and FPR are calculated and plotted on the graph. The ROC curve is shown in Figure 3.1.

When comparing the performance of different classification models, if the ROC curve of one model is completely enclosed by the curve of another one, the latter is considered to be better than the former. If the ROC curves of the two models cross, we generally compare the area under ROC curve (AUC), and the model with higher AUC is considered to be better (Hanley and McNeil, 1983). AUC mainly reflects the ranking quality of sample prediction, which refers to the probability that if a positive sample and a negative sample are randomly

selected, the model judges that the predicted value of the positive sample is higher than the predicted value of the negative sample. If AUC is equal to 1, the classification effect of the model is the best.

Because ROC curve shows the classification performances under different thresholds, there is an optimal threshold among them. This threshold corresponds to the point closest to the upper left corner of the ROC curve, at which point the difference between TPR and FPR reaches the maximum. In the following, this thesis will try to find the optimal classification threshold in each model, and then use the performance under the optimal threshold as the criterion for classifier selection.

3.2 Feature processing methods

3.2.1 Principal component analysis

Principal Component Analysis (PCA) refers to transforming a group of potentially correlated variables into a group of linearly uncorrelated variables through orthogonal transformation. The transformed variables are called principal components, and the number one principal component contains the most information. Through PCA, the information of high-dimensional data can be refined into several principal components, which can then be visualized in low-dimensional space, and graphical methods can be used to observe whether the data set is linearly separable.

Assuming that the data set $D = \{x_1, x_2, \dots, x_p\}$, contains p vectors, where $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$, each feature has m dimensions. The specific process of reducing the feature dimensions to d through PCA is as follows.

First, centralize processing of all samples: $x_i \leftarrow x_i - \frac{1}{p} \sum_{i=1}^p x_i$;

Second, calculate the covariance matrix of the sample XX^T ;

Third, perform eigenvalue decomposition on the covariance matrix XX^T , and take the eigenvectors corresponding to the largest d eigenvalues $a_1, a_2, \dots, a_d$;

Fourth, normalize all a_i , so that $a_i a_i^T = 1$;

Fifth, generate new features after dimensionality reduction: $Y = Ax_i, i = 1, \dots, d$, where $A = [a_1, a_2, \dots, a_d]^T, a_i = (a_{i1}, a_{i2}, \dots, a_{im})$.

3.2.2 Feature selection

Feature selection is an important process of data preprocessing. Its purpose is to remove irrelevant features, and reduce feature dimensions, which is similar to PCA. The difference is that after PCA dimensionality reduction, each new principal component no longer has the original feature meaning, and the variable after feature selection still maintains the original feature meaning.

There are three types of feature selection methods commonly used: filtering method, wrapping method and embedded method. Among them, the filtering method is independent of the classification method to process the data, that is, it removes unnecessary features before the model training, and then uses the learning algorithm to generate a classifier. It includes zero-variance filtering, chi-square filtering (Kira & Rendell, 2014) and F Filtration (Kononenko, 1994). Wrapped filtering refers to that, given a classifier, we use the performance of the classifier as a basis to select a subset of features. The feature selection is not independent of the classification method. This method requires multiple training of the classifier, and the computational cost is relatively high. Different from the previous two methods, the embedded method combines the feature selection with the classifier training, both are completed in the same optimization process, and feature selection is automatically performed during the classifier training process. Due to the large amount of data in the field of data mining, the filtering methods are commonly used. In the following, this thesis also mainly uses the filtering method to carry out the preliminary screening of features.

3.3 Classification algorithms

3.3.1 Support Vector Machine

Support Vector Machine (SVM) is a kind of small sample learning method with a solid theoretical foundation, which can be used to deal with linear and nonlinear classification problems without making assumptions about the distribution of data. The basic idea of SVM is to find a hyperplane $w^T x + b = 0$ in the sample space based on the training set T . The samples on one side of the hyperplane are classified as -1, and the samples on the other side are classified as +1. When the data is linearly separable, a method for solving the maximum interval is given. While when the data is non-linearly separable, the nonlinear problem in the low-dimensional space is transformed into the linear problem in the high-dimensional space through nonlinear mapping.

3.3.1.1 Linear SVM

Given a training sample set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$, where $x_i \in \mathbb{R}^n$ is the feature vector and $y_i \in \{1, -1\}$ is the label of x_i . The classic SVM is to find a hyperplane with the largest interval and the least classification error, as shown in Figure B1.

The basic mathematical model is shown in formula (3.8).

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^l h(w, b; x_i, y_i) \quad (3.8)$$

Among them, $w \in \mathbb{R}^n, b \in \mathbb{R}$, the first term $\frac{1}{2} \|w\|_2^2$ is the objective function that maximizes the interval in Figure B1; the second term is the hinge loss function, where $h(w, b; x_i, y_i) = \max\{1 - y_i(w^T x_i + b), 0\}$, which is the punishment to samples that are classified incorrectly (the samples in the interval in Figure B1). Parameter C is a penalty coefficient that adjusts the proportional relationship between the two items. The larger the C , the greater the penalty for misclassification.

After introducing the slack variable $\xi \in \mathbb{R}^l$, the above formula can be rewritten as formula (3.9).

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^l \xi_i \\ \text{s.t.} \quad & y_i (w^T x_i + b) \geq 1 - \xi_i, \forall i; \\ & \xi_i \geq 0, \forall i. \end{aligned} \quad (3.9)$$

Using the Lagrange multiplier method, we get the function (3.10).

$$L(w, b, \xi; \alpha, \theta) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^l \xi_i + \sum_{i=1}^l \alpha_i (1 - \xi_i - y_i (w^T x_i + b)) - \sum_{i=1}^l \theta_i \xi_i \quad (3.10)$$

Among them, $\alpha_i \geq 0, \theta_i \geq 0$ are the Lagrangian multipliers corresponding to the constraints. The partial derivatives of w, b, ξ are respectively obtained and set to be equal to 0, so that there are the following first-order conditions.

$$w = \sum_{i=1}^l \alpha_i y_i x_i \quad (3.11)$$

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad (3.12)$$

$$\alpha_i + \theta_i = C, \quad \forall i \quad (3.13)$$

Substituting formulas (3.11), (3.12), (3.13) into (3.10), the dual problem of SVM can be obtained.

$$\begin{aligned}
 & \min_{\alpha} \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i,j=1}^l \alpha_i \\
 & \text{s.t.} \quad y^T \alpha = 0; \\
 & \quad 0 \leq \alpha_i \leq C, \forall i.
 \end{aligned} \tag{3.14}$$

By solving the above problems, the optimal solution $\alpha^* = (\alpha_1^*, \dots, \alpha_i^*, \dots, \alpha_l^*)$ can be obtained, where α_i^* represents the optimal solution of α^i . As shown below, the classification threshold b^* is calculated, and finally the discriminant model of SVM is obtained.

$$y = \text{sgn}\left\{\sum_{i=1}^l \alpha_i^* y_i (x_i x^T) + b^*\right\} \tag{3.15}$$

$$b^* = y_i - \sum_{i=1}^l \alpha_i^* y_i (x_i x_j) \tag{3.16}$$

When entering the feature data of testing sample into the formula (3.15), if the value of $\sum_{i=1}^l \alpha_i^* y_i (x_i x^T) + b^*$ is positive, the predicted y of the sample is +1 and it is classified as positive, otherwise it is classified as negative.

3.3.1.2 Non-linear SVM

For the linearly inseparable data in the low-dimensional space, that is, when the classification effect of linear SVM is poor, the data can be mapped to the high-dimensional space through a nonlinear transformation $\phi(\cdot)$, to make it linearly separable in the high-dimensional space. The principle is shown in Figure B2.

Let $\phi(x)$ denote the feature vector after the sample x is mapped. The new feature space has a very high dimensionality, even infinite dimensionality, so the kernel function is introduced.

$$K(x_i, x_j) = (f(x_i), f(x_j)) \tag{3.17}$$

Then the formula (3.14) can be written as follows.

$$\begin{aligned}
 & \min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - 1_l^T \alpha \\
 & \text{s.t.} \quad y^T \alpha = 0; \\
 & \quad 0 \leq \alpha \leq C 1_l.
 \end{aligned} \tag{3.18}$$

Among them, the matrix $Q = \text{Diag}(y) K \text{Diag}(y) \in R^{l \times l}$, $K \in R^{l \times l}$, its elements $K_{ij} = K(x_i, x_j)$, $K(\cdot)$ is the kernel function, $\text{Diag}(y)$ is a square matrix with diagonal elements y and other elements 0, and 1_l is a vector composed of l ones.

When a new sample x is given, the following discriminant function can be used to judge whether it is positive or negative.

$$\begin{aligned}
y &= \text{sgn}(w^T \phi(x) + b) \\
&= \text{sgn}\left\{\sum_{i=1}^l \alpha_i y_i \phi(x_i)^T \phi(x) + b\right\} \\
&= \text{sgn}\left\{\sum_{i=1}^l \alpha_i y_i K(x, x_i) + b\right\}
\end{aligned} \tag{3.19}$$

It can be seen from the above process that, by using the kernel function $K(\cdot)$, it is not necessary to calculate the inner product of the nonlinear transformation $\phi(\cdot)$ in the high-dimensional space, but the dimension increase of the data and the linear classification in the high-dimensional space can be realized. Some commonly used kernel functions are shown in Table A3.

Keerthi and Lin (2003) pointed out that the linear kernel is a special case of Radial Basis Function kernel (RBF), H.-T. Lin and Lin (2003) found that the performance of the Sigmoid kernel is similar to that of RBF, and A.-Y. Chang et al. (2001) recommend RBF to be the first choice in general for the following reasons. First, after the RBF kernel maps the finite-dimensional data to the infinite-dimensional space, the mapped sample data is linearly independent, ensuring that a linear hyperplane can be found to classify the samples. Second, after the RBF kernel maps the sample data, the distance between the sample points in the high-dimensional space is proportional to the distance in the original space, the topological structure of the original data remains unchanged, and the information is not lost. Therefore, in the subsequent chapters of this thesis, the RBF kernel is used as the kernel function of SVM to map low-dimensional data to high-dimensional data.

3.3.2 Random Forest

Random forest (RF) is a kind of ensemble learning, which refers to the completion of learning tasks by constructing and combining multiple learners. The general structure is as follows. First generate a group of individual learners, and then use a certain strategy to combine them. The individual learner is usually generated from training data by an existing learning algorithm, such as Decision Tree, BP Neural Network. According to the generation method of the individual learner, the ensemble learning methods can be roughly divided into two types. The first type is a kind of serialization method. There is a strong dependency between individual learners and the learners must be generated serially. The representative is Boosting. The second type is a kind of parallelization method. There is no strong dependency between individual learners and the learners can be generated at the same time. The representatives are Bagging (Breiman, 1996) and RF (Breiman, 2001).

Based on the bootstrapping method Breiman, 1996 proposed, Bagging performs random sampling with replacement on a given data set containing m samples. After m sampling, a sample set containing m samples will be obtained. The purpose of using overlapping feature sets is to improve the independence of individual learners while ensuring that the performance of each individual learner is not too bad. Repeat this T times to obtain T sample sets containing m training samples, and train a base learner based on each sample set, and then combine these learners.

RF is an extended variant of Bagging. As shown in Figure B3, it builds Bagging integration based on the decision tree learner, and further introduces random feature selection in its training process. Specifically, the traditional decision tree selects an optimal feature from the feature set of the current node (assuming there are d features), while for each node of the base decision tree, RF first randomly selects a subset of k features from the feature set of the node, and then selects an optimal feature from this subset for division. The parameter k here controls the degree of randomness. If $k = d$, the construction of the base decision tree is the same as the traditional decision tree, while if $k = 1$, a feature is randomly selected for division. In general, $k = \log_2 d$ (Breiman, 2001).

Because of its simple logic but extremely strong performance, RF is known as a method that represents the level of ensemble learning technology. Unlike Bagging which only uses sample perturbation, RF uses both sample perturbation and feature perturbation to increase the differentiation of the base learners, to ensure the generalization performance of the ensemble learner.

3.3.2.1 Basic principle of Decision Tree

A decision tree contains a root node, several internal nodes and several leaf nodes. The leaf node corresponds to the decision result, and each other node corresponds to a feature test. The sample set contained in each node is divided into sub-nodes based on the result of the feature test. The root node contains the complete set of samples. The path from the root node to each leaf node corresponds to a decision test sequence. How to select the optimal feature of the root node and each internal node? Generally speaking, as the division process continues, we hope that the samples contained in the branch nodes of the decision tree belong to the same category as much as possible, that is, the purity of the nodes is getting higher and higher. There are four main indicators for measuring purity: information entropy, information gain (J. Ross Quinlan, 1986), information gain rate (J Ross Quinlan, 1993) and Gini index (Breiman et al., 2017).

Information entropy is the most commonly used indicator to measure the purity of a sample

set. Assuming that the proportion of the k -th sample in the current sample set D is p_k ($k = 1, 2, \dots, K$), then the information entropy of D is defined as shown in formula (3.20).

$$Ent(D) = - \sum_{k=1}^K p_k \log_2 p_k \quad (3.20)$$

Assuming that the discrete feature a has V possible values $\{a^1, a^2, \dots, a^V\}$, if feature a is used to divide the sample set D , V branch nodes will be generated. D^v refers to the number of samples whose value of feature a is a^v on the v -th branch node. Based on this, the information gain can be defined as formula (3.21). Generally speaking, the greater the information gain, the greater the purity improvement obtained by using feature a for division.

$$Gain(D, a) = Ent(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} Ent(D^v) \quad (3.21)$$

The information gain criterion prefers features with a large number of possible values. In order to reduce the possible adverse effects of this preference, J Ross Quinlan (1993) proposed not to directly use information gain in the C4.5 decision tree, but to use gain ratio to select the optimal feature, the formula is shown in (3-22), where IV is the intrinsic value of feature a J Ross Quinlan (1993), which is proportional to the number of values of a .

$$Gain_ratio(D, a) = \frac{Gain(D, a)}{IV(a)} = \frac{Gain(D, a)}{- \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \sum_{v=1}^V \frac{|D^v|}{|D|}} \quad (3.22)$$

The CART decision tree (Breiman et al., 2017) uses the Gini index to select the division features, and the purity of the data set D can be measured by the Gini value, as shown in formula (3.23). The Gini index of attribute a is defined as formula (3.24).

$$Gini(D) = \sum_{k=1}^K \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^K p_k^2 \quad (3.23)$$

$$Gini(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} Gini(D^v) \quad (3.24)$$

3.3.2.2 Feature importance evaluation

In addition to being used for classification, RF can also return the importance of each feature, which is determined by the average contribution of each feature to each decision tree and can be used for feature screening. Generally, Gini index or out-of-bag data error rate can be used to evaluate the contribution of features to the tree. The following will introduce the feature importance evaluation method based on the Gini index.

Assuming there are C features, $VIM_j^{(Gini)}$ represents the feature importance of the j -th

feature X_j , which refers to the average Gini index score of the node split impurity of the j -th feature in all decision trees of the random forest. The formula of Gini index is shown in formula (3.25).

$$GI_m = \sum_{k=1}^K \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^K p_{mk}^2 \quad (3.25)$$

Among them, K represents the total number of categories, and p_{mk} represents the proportion of the k -th category in node m . The importance of feature X_j at node m is measured by the change in the Gini index before and after the branch of node m .

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \quad (3.26)$$

GI_l and GI_r refer to the Gini indices of two new nodes branched from node m . If the node of feature X_j in the i -th decision tree is in set M , then the importance of feature X_j in the i -th decision tree can be expressed as formula (3.27).

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \quad (3.27)$$

If there are n trees in the random forest, the importance of the feature X_j in the entire random forest can be expressed as formula (3.28).

$$VIM_j^{(Gini)} = \sum_{i=1}^n VIM_{ij}^{(Gini)} \quad (3.28)$$

Finally, we normalize the importance scores of all features, and the value of the importance of feature X_j is obtained.

$$VIM_j = \frac{VIM_j^{(Gini)}}{\sum_{i=1}^C VIM_i} \quad (3.29)$$

3.3.3 Logistic Regression

Logistic regression (LR), also known as log probability regression, is a kind of linear regression classifier. LR inputs the linear regression result into the sigmoid activation function, converts the continuous result variable into a probability value in the interval $[0,1]$, and converts the probability value into a classification result output according to the probability threshold. The probability predictive function of LR is shown in formula (3.30).

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \quad (3.30)$$

Among them, θ is the parameter to be estimated, and the above formula can also be

changed to formula (3.31).

$$\ln \frac{h_{\theta}(x)}{1-h_{\theta}(x)} = \theta^T x \quad (3.31)$$

Given the feature x , $h_{\theta}(x)$ refers to the probability that the sample is positive, $1 - h_{\theta}(x)$ refers to the probability that the sample is negative, and the ratio of the two is called odds, which means the relative probability that x is positive. After calculating $h_{\theta}(x)$, the classification criteria of LR are as follows.

When $h_{\theta}(x)$ is greater than 0.5, the predicted label of the sample is +1 and classified as a positive category, which is referred to as a “will default” corporation in this thesis. When $h_{\theta}(x)$ is smaller than 0.5, the predicted label of the sample is -1 and classified as a negative category, which is referred to as a “will not default” corporation in this thesis.

How to choose the optimal value of the parameter θ ? This thesis uses the loss function, which is the performance of a certain set of parameter values on the training set, to measure the pros and cons of the parameters θ . The loss function is derived from the maximum likelihood method, and the result is shown in formula (3.32).

$$J(\theta) = -\sum_{i=1}^m (y_i * \log(h_{\theta}(x_i)) + (1 - y_i) * \log(1 - h_{\theta}(x_i))) \quad (3.32)$$

Among them, m is the number of samples in the training set, y_i is the true label of sample i , and $h_{\theta}(x_i)$ is the return value of sample i in the logistic regression model with parameter θ . In order to avoid over-fitting, a penalty term, also known as a regular term, is generally added to the loss function. The regular term is a multiple of the paradigm of the parameter vector θ . There are two types of paradigms: L1 paradigm and L2 paradigm. The specific expressions are as follows.

$$J(\theta)_{L1} = J(\theta) + \frac{1}{C} \sum_{j=1}^n |\theta_j| \quad (j \geq 1) \quad (3.33)$$

$$J(\theta)_{L2} = J(\theta) + \frac{1}{C} \sqrt{\sum_{j=1}^n (\theta_j)^2} \quad (j \geq 1) \quad (3.34)$$

In the above formulas, the parameter C is a hyper parameter used to control the degree of regularization, and n is the total number of features. Empirical evidence has shown that although both L1 and L2 can control over-fitting, as C increases, L1 regularization will compress the parameters to 0, and L2 regularization will make the parameters smaller, but not 0. Since this thesis plans to use RF for feature screening, when using LR to train the filtered features, L2 regularization will be used to keep the contribution of each feature to the classification result

as much as possible.

3.3.4 BP Neural Network

BP Neural Network (BPNN) refers to the Error Back Propagation neural network, which is a multi-layer neural network, including an input layer, a hidden layer and an output layer. Each layer contains multiple neuron nodes. The neurons in each layer are connected to each other, and the neurons in the same layer are independent of each other.

The principle of BPNN is as follows. Given the training set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, $x_i \in R^d$ is the input vector of the neural network, described by d features, and $y_i \in R^l$ is the output vector of the neural network, which is an l -dimensional real-valued vector. Figure B4 shows a multi-layer BPNN structure with d input neurons, l output neurons, and q hidden layer neurons. The threshold of the j -th neuron in the output layer is represented by θ_j , and the threshold of the h -th neuron in the hidden layer is represented by γ_h . We take the h -th neuron in the hidden layer as the observation object. The input signal received by this neuron is $\alpha_h = \sum_{i=1}^d v_{ih} x_d$, and the output signal is b_h , which is derived from the input signal through sigmoid transformation, so there is $b_h = f(\alpha_h - \gamma_h) = 1/(1 + \exp(-(\alpha_h - \gamma_h)))$. The signal received by the j -th neuron in the output layer is $\beta_j = \sum_{h=1}^q w_{hj} b_h$, and the output value is $\hat{y}_j = f(\beta_j - \theta_j)$. Based on the above definitions, for the sample (x_k, y_k) in the training set, after inputting its feature x_k into the input layer, we will obtain $\hat{y}_k = (\hat{y}_1^k, \hat{y}_2^k, \dots, \hat{y}_l^k)$ in the output layer, the mean square error of the neural network on (x_k, y_k) is $E_k = \frac{1}{2} \sum_{j=1}^l (\hat{y}_j^k - y_j^k)^2$.

The training process of BPNN is mainly divided into two stages. In the first stage, the signal propagates forward. After determining the initial values of the above-mentioned weights v and w and the threshold parameters γ and θ , the signal enters the input layer, is transformed by the Sigmoid function of the hidden layer, and the value of the output layer is obtained. In the second stage, the error is propagated back. It is transmitted from the output layer to the hidden layer, and finally to the input layer, w , θ , v and γ are adjusted in turn. Then we restart a new round of signal forward propagation and error back propagation process, and continue until the number of learning reaches the set value or the error is reduced to the required level. The parameters are adjusted in the negative gradient direction of the objective function, according to the gradient descent method. Under the premise of a given learning rate η , the adjustment formulas for the parameters are shown in (3-35) -(3-38).

$$w_{hj} = -\eta \frac{\partial E_K}{\partial w_{hj}} = \eta \hat{y}_j^k (1 - \hat{y}_j^k) (y_j^k - \hat{y}_j^k) b_h \quad (3.35)$$

$$\theta_j = -\eta \hat{y}_j^k (1 - \hat{y}_j^k) (y_j^k - \hat{y}_j^k) \quad (3.36)$$

$$v_{ih} = \eta b_h (1 - b_h) x_i \sum_{j=1}^l w_{hj} \hat{y}_j^k (1 - \hat{y}_j^k) (y_j^k - \hat{y}_j^k) \quad (3.37)$$

$$\gamma_h = -\eta b_h (1 - b_h) \sum_{j=1}^l w_{hj} \hat{y}_j^k (1 - \hat{y}_j^k) (y_j^k - \hat{y}_j^k) \quad (3.38)$$

3.3.5 Transductive Support Vector Machine

Transductive Support Vector Machine (TSVM) is the most famous type of semi-supervised SVM (Joachims, 1999). This method is similar to the classic SVM, and both are learning methods for binary classification problems. The difference between the two is: SVM is a supervised algorithm that only uses labeled samples to train the classifier, while TSVM is a semi-supervised algorithm that uses both labeled samples and unlabeled samples to train the classifier. In the question of this thesis, we can collect the information of those companies that have passed the loan approval and those that have not passed the approval. For companies that have passed loan approval, we can observe their “default” or “non-default” behaviors and get their default label, while the default label of companies that have not passed cannot be observed. Therefore, this thesis can collect a large number of unlabeled samples, and TSVM algorithm can effectively use this data, which may help to improve the model’s performance in predicting the risk of corporate default.

The principle of the TSVM algorithm is as follows. Given labeled samples $D_l = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ and unlabeled samples $D_u = \{x_{l+1}, x_{l+2}, \dots, x_{l+u}\}$, where $y_i \in \{-1, +1\}$, $l \ll u$, $l + u = m$. The purpose of TSVM is to give prediction labels $\hat{y} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \dots, \hat{y}_{l+u})$, $\hat{y}_i \in \{-1, +1\}$ of the samples in D_u . The object and the constraints are shown in formula (3.39).

$$\begin{aligned} \min_{w, b, \hat{y}, \xi} \quad & \frac{1}{2} \|w\|_2^2 + C_l \sum_{i=1}^l \xi_i + C_u \sum_{i=l+1}^m \xi_i \\ \text{s.t.} \quad & y_i (w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, l; \\ & \hat{y}_i (w^T x_i + b) \geq 1 - \xi_i, i = l+1, l+2, \dots, m; \\ & \xi_i \geq 0, \forall i. \end{aligned} \quad (3.39)$$

Among them, ξ is the slack vector, and C_l and C_u are the compromise parameters used to balance the importance of labeled and unlabeled samples in the model. Its workflow is as

follows.

First, use the labeled data D_l to train an SVM learner, use the learner to predict the unlabeled sample D_u , and get $\hat{y} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \dots, \hat{y}_{l+u})$;

Second, set the initial values of C_l and C_u . Make $C_u \ll C_l$ to ensure that the unlabeled samples have less impact on classification, because unlabeled samples have a higher probability of misclassification in the model trained earlier;

Third, find two unlabeled samples with opposite labels and the sum of the slack variables is greater than 2, exchange their labels, and solve (3.39) again. In the process, let $C_u = \min \{2C_u, C_l\}$ to keep increasing the value of C_u , stop the loop until $C_u = C_l$ and the optimal parameters are solved. The last classifier is used to predict the unlabeled samples, and the testing set is used to test the performance of the classification model.

Chapter 4: Analysis on the Credit Evaluation Index System of “Tax E Financing”

Constructing a reasonable credit evaluation index system is the basis for credit evaluation of “Tax E Financing”, and it is also one of the key links. The rationality of the index system and the quality of the feature data determine the upper limit of the classification model, and better algorithms and better parameters just keep approaching this upper limit. In the existing literature, the credit evaluation indicators for small, medium and micro enterprises include financial indicators and non-financial indicators. Among them, financial indicators include solvency, operating ability, profitability and development ability; non-financial indicators include business owner characteristics, company development tendency, strategic planning, market conditions, growth, supply chain advantages, and other risk indicators. Non-financial information occupies a huge proportion in the evaluation index system. The above index system mainly has the following problems. First, the cost of obtaining data is high. Collecting the non-financial information requires a lot of time and represents a high cost for the bank. Second, the reliability and completeness of the financial statements of small, medium and micro enterprises is low. As a result, their financial information not only fails to correctly evaluate the credit risk of the enterprise, but may become information noise. Third, it is difficult to quantify qualitative indicators, and subjective judgments in this process will also increase information noise.

In order to accurately identify the credit risk of small, medium and micro enterprises, reduce the time and cost of information collection, avoid subjectivity in the quantification of qualitative indicators, and improve the integrity and reliability of indicator data, we in this thesis believe that the tax information provided by the government taxation department should be introduced into the credit evaluation index system of small, medium and micro enterprises. The reason is as follows. First, the cost of obtaining tax information is relatively low. In the context of “Interaction between Banks and Tax Authorities”, local government tax authorities will share the tax information of loan applicant enterprises with banks free of charge. Second, the tax information is accurate. Different from the financial characteristics, enterprise development, strategy and other information provided by the enterprise itself, there may be cases where the data is whitewashed in order to successfully obtain loans. Tax data is the information formed by the enterprise’s true tax payment and declaration behavior, which is absolutely accurate and

true. Third, tax information is highly correlated with the true financial status of the enterprise and can reflect its operating performance.

This chapter adds tax characteristics to the existing credit evaluation index system for small, medium and micro enterprises, and collects all the available characteristic data of defaulting enterprises, non-defaulting enterprises and non-passing enterprises from the three aspects of basic information, financial information and tax information. Based on the collected sample data, this chapter will profile the defaulting enterprises, initially explore the characteristics that can effectively classify defaulting enterprises and non-defaulting enterprises, analyze the characteristics of non-passing enterprises to further judge the possibility of their misclassification.

4.1 Sample data collection and processing

4.1.1 Sample data collection

This thesis will collect the characteristic data of non-defaulting enterprises, defaulting enterprises and non-passing enterprises. In order to reduce the impact of external environment change on the enterprise operation, this thesis uniformly selects enterprises which have applied for loans through the “Tax E Financing” of Nantong Branch of Jiangsu Bank in the second half of 2017, and ensures that each enterprise has complete basic information, including loan application amount, loan application date, number of employees, registered capital, business registration type, industry and business opening time; complete financial information from 2016 to 2018, including balance sheet and income statement; complete tax information, including tax credit rating, taxpayer type, number of tax violations in the 36 months before loan application, value-added tax and income tax payment amount from 2016 to 2018. In addition, since the personal information of enterprise owners used by banks for loan review comes from the short-term authorization of the People’s Bank of China, the information cannot be saved and downloaded from the “Tax E Financing” system, so it cannot be obtained for this thesis. We matched the basic information, financial information, and tax information of every enterprise, and eliminated sample enterprises with large missing information. The complete information of 1,304 non-passing enterprises, 546 non-defaulting enterprises, and 13 defaulting enterprises were collected.

Because the size of defaulting sample is too small, it will seriously reduce the classification accuracy of the model for defaulting enterprises. In order to obtain more defaulting sample, this thesis supplements the collection of information on defaulting enterprises that applied for loans

through the “Tax E Financing” in the second half of 2018 and the full year of 2019. Among them, 18 enterprises applied for loans in 2018 and 23 in 2019. We have obtained information on 54 defaulting enterprises in total.

4.1.2 Generating new features

4.1.2.1 Basic features

(1) Industry

According to the original data collected, the industry distribution of the three samples is shown in Table 4.1. It can be seen that most of the enterprises applying for loans are from manufacturing and wholesale and retail industries, while there are relatively few companies in other industries. Moreover, the industry features of the non-defaulting sample and defaulting sample are inconsistent with the non-passing sample, which means that when we use the classifiers trained with defaulting and non-defaulting enterprises, we cannot effectively classify the current non-passing enterprises because new industries that do not exist in the classifier appear in the non-passing enterprises. Therefore, in order to ensure that the industry feature of training data, testing data, and predicting data are consistent, this thesis reprocesses the industry feature, and merges all industries except manufacturing, wholesale and retail into “other industries”, reducing the current 14 kinds of industries to 3 kinds.

Table 4.1 Industry distribution of different samples

Industry	Non-passing	Defaulting	Non-defaulting
Manufacturing	672	29	323
Wholesale and retail	435	16	174
Construction	57	5	22
Transportation, storage and postal	37	0	10
Leasing and business services	29	0	0
Scientific research and technical service	28	1	5
Accommodation and catering	22	2	4
Resident services, repairs and other services	11	0	7
Real estate	5	0	0
Agriculture, forestry, animal husbandry and fishery	4	0	0
Culture, sports and entertainment	1	1	0
Finance	1	0	0
Electricity, heat, gas and water production and supply	1	0	1
Water conservancy, environment and public facilities management	1	0	0
Total	1304	54	546

Source: Author.

(2) Enterprise's duration

The original data collected in this thesis includes the time of business opening and the time that the company applied for a loan through “Tax E Financing”. This thesis uses the loan application year minus business opening year to generate the variable of the enterprise’s duration. The larger the value of this variable, the longer the operating time of the enterprise, and the more stable the operating conditions may be.

4.1.2.2 Financial features

The original financial data of all samples collected in this thesis is the balance sheet and income statement of the financial year before the loan application. These financial data are total data, such as total assets and total profits, which are not comparable between enterprises of different sizes. In order to better reflect the financial status of the enterprise and eliminate the scale effect, this thesis calculates a series of ratio indicators, which are used to measure the solvency, profitability and operating capacity of the enterprise. The specific indicators and calculation formulas are shown in Table 4.2. Among them, the lower the gearing ratio, the stronger the enterprise’s long-term debt solvency; the higher the current ratio, the quick ratio, and the cash ratio, the stronger the enterprise’s short-term debt solvency; the higher the sales margin, the return on assets (ROA), and the return on equity (ROE), the higher the enterprise’s profitability; the higher the total asset turnover rate, inventory turnover rate and accounts receivable turnover rate, the stronger the enterprise’s asset management capabilities. At the same time, the stronger the inventory turnover rate and the accounts receivable turnover rate, the shorter the business cycle of the enterprise and the stronger the short-term debt solvency.

Table 4.2 Added financial features and calculation formulas

Financial condition	Index	Calculation formulas
Solvency	Gearing ratio	Total liabilities/Total assets
	Current ratio	Current assets/Current liabilities
	Quick ratio	Quick assets/Current liabilities
	Cash ratio	Cash and equivalents/Current liabilities
Profitability	Sales margin	Net profit/Operating income
	Return on assets	Net profit/Total assets
	Return on equity	Net profit/Equity
	Total assets turnover	Operating income/Total assets
Operational capability	Inventory turnover	Operating cost/Inventories
	Accounts receivable turnover	Operating income / Accounts receivable

Source: Author.

4.1.2.3 Tax features

(1) Tax credit rating

Table A4 reflects the tax credit rating of different samples. There are the same problems as the industry feature, that is, there are new features in the non-passing sample that the defaulting

sample and the non-defaulting sample do not have. For example, the tax credit rating of some non-passing enterprises is D. Therefore, in order to make the features of all samples consistent, and considering that the enterprises with tax credit ratings of A and B account for the highest proportion, therefore in this thesis we classified both C and D into B, reflecting the lower tax credit rating. Therefore, the Tax credit ratings change from four to two.

(2) Tax amount and frequency

In the original data collected for this thesis, the tax data of most enterprises is the monthly value-added tax (VAT) payment amount and the quarterly income tax payment amount from 2016 to 2018. The defaulting enterprises which applied for loans in 2018 and 2019 have monthly VAT payment data and quarterly income tax payment data for 2018 and 2019. Because different enterprises apply for loans at different times, in order to accurately reflect the tax status before the loan application, in this thesis we extract the monthly income VAT payment data and quarterly income tax payment data for the half year before the loan application through “Tax E Financing”. Then, for this thesis we calculate the total income tax and VAT payment amount in this period. The higher the tax amount, the better the business performance of the enterprise.

For this thesis we also calculated the frequency of income tax and VAT payment made by each enterprise for the half year before applying for the loan. The more the number of tax payments, the more stable the business performance of the enterprise. Meanwhile, according to the tax violation information of each enterprise, the number of tax violations in the first 6 months, 12 months, 24 months and 36 months before the loan application was also calculated.

After generating these new features, the data set collected for this thesis has a total of 1904 enterprises, of which 546 are non-defaulting, 54 are defaulting, and 1304 are non-passing. The number of features is 81, including 6 basic features, 56 financial features and 19 tax features.

4.2 Feature analysis of sample enterprises

4.2.1 Analysis on the features of defaulting enterprises

This thesis first compares the features of defaulting enterprises and non-defaulting enterprises, and portrays the defaulting ones. For continuous feature variables, this thesis conducts a significance test on the difference between the defaulting and non-defaulting enterprises. Because the defaulting sample and non-defaulting sample have different sizes, we adopt the Mann-Whitney U test, which was proposed by H.B. Mann and D.R. Whitney in 1947. It assumes that the two samples are from two populations that are exactly the same except for the

population mean, and the purpose is to test whether the means of the two populations are significantly different.

(1) Differences in basic features

Table A5 shows the differences between defaulting and non-defaulting enterprises in terms of continuous basic feature variables. As shown in the table, there are two basic features that have significant differences between defaulting sample and non-defaulting sample, namely, the loan application amount and the enterprise's duration. The average loan application amount of the non-defaulting sample is 1.5652 million RMB, and that of the defaulting sample is 0.67 million RMB. The difference between the two is 0.894 million RMB, which is significant at the 1% level. In the non-defaulting sample, the average duration is 9.03 years, while that of defaulting sample is 8.06 years. The difference between the two is 0.97 years, which is significantly different from 0 at the level of 10%. These results show that the loan application amount of defaulting enterprises is relatively low and the duration is relatively short.

Meanwhile, it can be seen that the number of employees of defaulting sample is relatively small and the registered capital is relatively high, but the statistical difference between the two is not significant.

For the non-continuous basic feature variables, this thesis calculates their default distribution. Table A6 shows the default rate of different business registration type. As shown in the table, among the 600 enterprises that have passed the loan approval, the default rate of private limited company is 11.29%, which is the highest among the three types, and the second is other limited company, the default rate of which is 5.63%. The proportions of the two in the total enterprises are 60.5% and 38.5% respectively.

Table A7 shows the default rate of different industries. It can be seen that enterprises in other industries have the highest default rate of 15.52%, while the default rates of manufacturing and wholesale and retail industries are relatively close, being 8.24% and 8.42% respectively.

(2) Differences in financial features

In this thesis we will compare the differences of the total amount of accounting items between defaulting sample and non-defaulting sample, as well as the differences in financial ratios. Among them, accounting items include balance sheet items and income statement items. Table 4.3 shows the difference between the defaulting and non-defaulting samples in each item of the balance sheet. Since the interest receivable of the two samples are all 0, they are deleted.

We can find that there is no significant difference in the total assets of defaulting sample and non-defaulting sample, but there is a significant difference in the asset structure. In terms of current assets, the default sample has significantly lower short-term investment, notes

receivable, dividends receivable, and other current assets, while the accounts receivable, prepayments, other receivables and inventories of defaulting sample are significantly higher than that of non-defaulting sample. In the part of non-current assets, the defaulting sample's long-term equity investment, accumulated depreciation, engineer materials and fixed assets in liquidation are significantly lower than those of non-defaulting sample, while construction in progress, intangible assets and long-term prepaid expenses are significantly higher than those of non-defaulting sample. Among them, the average value of construction in progress of defaulting sample is 36.24 times that of non-defaulting sample, and other receivables are 5.16 times that of non-defaulting sample.

Table 4.3 Balance sheets comparison of defaulting and non-defaulting samples

Features	Non-defaulting	Defaulting	Difference	P value
Cash	775882	975231	-199349	0.335
Short-term investment	2661	0	2661	0.000***
Notes receivable	70349	10000	60349	0.000***
Accounts receivable	2473046	3217404	-744358	0.005***
Prepayments	338364	783695	-445331	0.000***
Dividends receivable	3523	0	3523	0.000***
Other receivables	623425	3214951	-2591526	0.000***
Inventories	952496	1271499	-319003	0.001***
Other current assets	2074	0	2074	0.000***
Total current assets	5251219	7980279	-2729059	0.102
Long-term equity investments	47624	39286	8338	0.000***
Fixed assets(original value)	2176117	1460052	716065	0.115
minus: accumulated depreciation	815111	755642	59468	0.001***
Engineer materials	2268	0	2268	0.000***
Construction in progress	50287	1822595	-1772308	0.000***
Fixed assets in liquidation	668	0	668	0.000***
Intangible assets	157372	328435	-171062	0.000***
Long-term prepaid expenses	47584	102596	-55011	0.000***
Total assets	6917595	9677358	-2759763	0.353
Short-term debt	886718	1215098	-328380	0.000***
Notes payable	19277	2366369	-2347091	0.000***
Accounts payable	1358177	3282626	-1924450	0.000***
Advance receipts	145301	447586	-302285	0.000***
Wages payable	92020	255284	-163264	0.000***
Other payables	1041912	1418571	-376659	0.000***
Other current liabilities	1272	57227	-55955	0.000***
Total current liabilities	3575804	5900316	-2324512	0.046**
Long-term debt	32275	48688	-16413	0.000***
Long-term payables	8529	712001	-703472	0.000***
Total Liabilities	3617596	6142763	-2525167	0.041**
Equity	3300825	2832493	468332	0.608

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

In terms of liabilities, the amount of various liabilities of defaulting sample is significantly higher than that of non-defaulting sample. Among them, the notes payable of the defaulting sample is 122.76 times of the non-defaulting one, other current liabilities are 44.98 times of the

non-defaulting one, and the long-term payables are 83.48 times of the non-defaulting one. In the part of shareholders' equity, there is no significant difference in the total equity of the two samples.

Table 4.4 shows the differences in each item of the income statement between the defaulting and non-defaulting samples. It can be found that in terms of main business, the financial expenses of defaulting sample are 1.16 times that of non-defaulting sample, and the difference between the two is significant. In terms of non-operating business, the non-operating income of the defaulting sample is 5.13 times that of the non-defaulting sample, and the non-operating expenditure is 2.63 times that of the non-defaulting sample, and the difference between the two is statistically significant.

Table 4.4 Income statement comparison of defaulting and non-defaulting samples

Features	Non-defaulting	Defaulting	Difference	P value
Operating income	8925035	11267493	-2342457	0.310
Operating cost	8048221	10086485	-2038264	0.359
Business tax and surcharges	30728	30984	-256	0.319
Financial expenses	58441	67564	-9122	0.003***
Non-operating income	19631	100616	-80985	0.000***
Non-operating expenditure	10343	27241	-16898	0.000***
Profit	105615	25913	79702	0.630
Income tax	21890	15167	6724	0.677
Net profit	83725	10746	72978	0.611

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

Table 4.5 shows the differences in financial ratios between the defaulting sample and the non-defaulting sample. It can be seen that the average gearing ratio of the defaulting sample is 53.36%, which is significantly higher than the 46.67% of non-defaulting sample, indicating that defaulting enterprises have poor long-term solvency. However, among the indicators for measuring short-term solvency, the current ratio, quick ratio and cash ratio of defaulting sample are all higher than those of non-defaulting sample. Among them, the quick ratio of the defaulting enterprises is 1947, and the quick ratio of the non-defaulting enterprises is 22.36. The difference between the two is significant at the level of 5%, which indicates that the defaulting sample has stronger short-term solvency, but this is contradictory to its default behavior.

In order to eliminate the influence of extreme values on the mean value, this thesis further compares the median and 75% quantile of the quick ratios of the two samples, and finds that the median quick ratio of the defaulting sample is 1.47, slightly lower than that of the non-defaulting sample. However, the 75% quantile of defaulting sample is 2.95, which is still higher than the 2.74 of non-defaulting sample. On the one hand, it shows that the distribution of quick

ratios of defaulting sample is extremely uneven. On the other hand, it shows that a large part of defaulting enterprises has stronger short-term solvency than non-defaulting enterprises, reflecting that their financial statement data may be unreliable.

Table 4.5 Financial ratios comparison of defaulting and non-defaulting samples

Features	Non-defaulting	Defaulting	Difference	P value
Gearing ratio	46.67%	53.36%	-6.68%	0.007***
Current ratio	25.65	1753.69	-1728.03	0.308
Quick ratio	22.36	1947.70	-1925.35	0.019**
Cash ratio	13.68	15.98	-2.30	0.871
Sales margin	0.02	0.00	0.02	0.330
Return on assets	0.05	0.03	0.02	0.965
Return on equity	0.01	-0.02	0.03	0.427
Total assets turnover	1.96	3.83	-1.87	0.960
Inventory turnover	464.65	96.36	368.29	0.979
Accounts receivable turnover	10.09	4.63	5.46	0.874

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

(3) Differences in tax features

Table 4.6 shows the differences in tax features between the defaulting sample and the non-defaulting sample. It shows that there is no significant difference in the number of tax violations between defaulting enterprises and non-defaulting enterprises. However, the tax amount and the number of tax payments of non-defaulting enterprises are significantly higher than those of defaulting enterprises. In the six months before the loan application, the average amount of income tax and VAT paid by defaulting enterprises are 1,894.02 RMB and 34,342.5 RMB, respectively, which are only 12.36% and 26.51% of the amount of non-defaulting enterprises. During this period, the average number of income tax and VAT payments of defaulting enterprises is 0.33 and 0.94, which is only 24.81% and 19.79% of the number of non-defaulting enterprises. Meanwhile, we can also find that the tax information closer to the time of loan application has a greater difference between defaulting sample and non-defaulting sample. For example, the average amount of income tax paid by defaulting enterprises in the quarter before the loan application is 694.78 RMB, which is only 8.43% of the amount of non-defaulting enterprises, and the average VAT amount of defaulting sample in the first month before the loan application is 5,405.61 RMB, which is 24.04% of the amount of non-defaulting sample.

Re-analyzing the income tax data showed in Table 4.4, it can be found that the income tax amount of the defaulting enterprises is 69% of the non-defaulting enterprises, which is much higher than the ratio of the income tax amount of defaulting enterprises to that of non-defaulting enterprises as reflected in Table 4.6. Since the tax information is obtained from the government taxation department, it has complete accuracy and authenticity, which further reflects that the financial data provided by the enterprise may be unreliable at least to a certain extent.

Table A8 shows the default rates of enterprises with different tax credit ratings. From the table, it can be seen that the default probability of enterprises with tax credit rating of A is 6.23%, and enterprises with tax credit rating of B have a default probability of 11.31%. It shows that enterprises with lower tax credit ratings are more likely to default.

Table 4.6 Tax features comparison of defaulting and non-defaulting samples

Features	Non-defaulting	Defaulting	Difference	P value
Number of tax violations in the past six months	0.16	0.41	-0.25	0.261
Number of tax violations in the past year	0.32	0.56	-0.24	0.554
Number of tax violations in the past two year	0.57	0.91	-0.34	0.282
Number of tax violations in the past three year	0.84	1.19	-0.34	0.956
Income tax amount in the first quarter before the loan application	8239.65	694.78	7544.87	0.000***
Income tax amount in the second quarter before the loan application	7086.67	1199.24	5887.43	0.000***
VAT amount in the first month before the loan application	22150.84	5405.61	16745.23	0.000***
VAT amount in the second month before the loan application	22838.12	5565.17	17272.96	0.000***
VAT amount in the third month before the loan application	22446.40	6336.70	16109.70	0.000***
VAT amount in the fourth month before the loan application	22285.23	3940.96	18344.27	0.000***
VAT amount in the fifth month before the loan application	20152.79	7120.76	13032.03	0.000***
VAT amount in the sixth month before the loan application	19661.67	5973.30	13688.38	0.000***
Number of income tax payments in the past six months	1.33	0.33	1.00	0.000***
Number of VAT payments in the past six months	4.75	0.94	3.81	0.000***
Total amount of income tax in the past six months	15326.32	1894.02	13432.30	0.000***
Total amount of VAT in the past six months	129535.06	34342.50	95192.56	0.000***

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

Table A9 shows the default rates of different taxpayers. It can be seen from the table that the default probability of small-scale taxpayers is 10%, while the default probability of general taxpayers is 8.98%, indicating that small-scale taxpayers are more likely to default.

Through the above analysis, in this thesis we find that, compared with non-defaulting enterprises, defaulting enterprises have the following characteristics:

First, in terms of basic features, defaulting enterprises have a lower loan application amount and a shorter duration, and private limited companies and other industries have higher default

probability.

Second, in terms of financial features, defaulting enterprises have higher accounts receivable, prepayments, other receivables and inventories, indicating that their accounts receivable occupy a large amount of funds and have a higher recovery risk, and the quality of their current assets is low. They have higher construction in progress, intangible assets and long-term prepaid expenses, indicating that their long-term investment efficiencies are low. Their gearing ratios are high, and all kinds of liabilities are relatively high, especially notes payable, other current liabilities and long-term payables, and their financial expenses are relatively high. They have higher non-operating income and profit. What is contradictory is that the defaulting enterprises have higher quick ratio, and their short-term debt repayment abilities are strong, which may be due to unreliable financial data.

Third, in terms of tax features, the amount and frequency of income tax and VAT paid by defaulting enterprises are significantly lower, and the tax information closer to the time of loan application has greater differences between defaulting sample and non-defaulting sample. Meanwhile, Enterprises with lower tax credit rating and small-scale taxpayers have a higher probability of default.

4.2.2 Analysis on the features of non-passing enterprises

As the approval rule of “Tax E Financing” is one-vote veto system, as long as the enterprise triggers the rejection rule, its loan application will be directly rejected. In this thesis we will further analyze the reasons for the rejection of the non-passing enterprises, and the differences between them and the defaulting and non-defaulting enterprises, to analyze the possibility of them being rejected by mistake.

4.2.2.1 Reasons for rejection of non-passing enterprises

Among the 1,304 non-passing enterprises, some enterprises trigger more than one rule. Table A10 shows the distribution of the number of rules that the non-passing enterprises trigger. 51.92% of non-passing enterprises trigger only 1 rule, 24.27% of non-passing enterprises trigger 2 rules, and 23.81% of them trigger 3 or more, which shows that most enterprises fail to pass loan approval because of one-sided reasons.

Table 4.7 shows the top ten reasons why the non-passing enterprises are rejected. It can be seen from the table that 16.71% of the non-passing enterprises are rejected because the gearing ratio exceeded 75%, 13.59% of them are rejected because the total tax payment in the past 12 months is less than 20,000, 7.5% of them are rejected because the tax credit rating is lower than

B, 6.52% of them are rejected because the actual income tax paid in the previous 24 months is less than 0, 6.44% of them are rejected because of the severe decline in the business, 6.04% of them are rejected because they had no tax declaration data in the past two months, 5.41% of them are rejected because their data was out of date, 5.09% of them are rejected because the total sales income of the last two periods is equal to 0, 3.44% of them are rejected because they are financial, construction or real estate enterprises, and 2.41% of these enterprises are rejected because their VAT has not been declared for three consecutive periods or six cumulative periods.

In summary, there are three types of reasons for loan rejection. The first type is related to the basic features of the enterprise, mainly considering the industry. The second category is related to the financial characteristics of the enterprise, mainly considering the gearing ratio, total income and its changes. The third category is related to the tax features of the enterprise, mainly considering the tax credit rating, tax amount, and the number of tax declarations. Through the analysis of 4.2.1, we can see that the above indicators are significantly different between non-defaulting enterprises and defaulting enterprises. However, it should be noted that the current loan review rule of “Tax E Financing” give each rule 100% weight, but in fact, the degree of influence of related characteristics on enterprise’s default behavior is likely to be different. At the same time, the current loan review rule also ignores many features that can effectively identify defaulting enterprises, such as the company’s duration, the most recent tax payment amount, and other information.

Table 4.7 Top 10 rules triggered and their percentages

Rule	Number of enterprises	Percentage	Rule description
R905	423	16.71%	For customers with financial statements, the ones that gearing ratio is more than 75% will be rejected, and no report does not trigger the rule.
R602	344	13.59%	Total tax payment <20,000 in the past 12 months.
R604	190	7.50%	Tax credit rating of the operating entity is not A or B.
R906	165	6.52%	Income tax actually paid in the previous 24 months is less than 0.
R115	163	6.44%	Business of the operating entity declines seriously.
R117	153	6.04%	There is no tax declaration data for the past two months.
Data expired	137	5.41%	Data expired
R910	129	5.09%	For customers who pay VAT (value added tax), the total sales income of the last two periods is equal to 0.
R908	87	3.44%	The operating entity is a financial, construction or real estate enterprise.
H910	61	2.41%	VAT has not been declared for three consecutive periods or six cumulative periods.

Source: Nantong Branch of Jiangsu Bank.

4.2.2.2 Features of non-passing enterprises

The analysis of 4.2.1 shows that there are significant differences in many characteristics of defaulting enterprises and non-defaulting enterprises. Considering that the financial information may have a certain degree of unreliability, in this thesis we will compare whether the basic features and tax features of the three types of samples of defaulting, non-defaulting and non-passing are also significantly different, and judge whether the non-passing enterprise is more inclined to defaulting or non-defaulting.

Table 4.8 shows the differences in the basic features of the three types of samples. It can be seen from the table that the average loan application amount and the number of employees of the non-passing enterprises are higher than those of the non-defaulting and the defaulting enterprises, and there is a significant difference from the non-defaulting enterprises. However, the duration of non-passing enterprises is close to that of defaulting ones, which is significantly lower than that of non-defaulting enterprises.

Table 4.8 Comparison of the basic features of the three types of samples

Features	Non-defaulting	Defaulting	Non-passing	Non-defaulting – non-passing	P value
Loan application amount (ten thousands)	156.524	67.093	161.126	-4.60	0.057*
Number of employees	15.840	12.370	19.992	-4.15	0.004***
Registered capital (ten thousands)	341.97	50.99	164.22	-130.02	0.235
Enterprise's duration	9.027	8.056	8.096	0.93	0.000***

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

Table 4.9 shows the differences in tax features of the three types of samples. It can be seen that the average number of tax violations of non-passing enterprises is significantly higher than that of non-defaulting and defaulting enterprises. However, the numbers of income tax and VAT payment of non-passing sample in the past six months are higher than those of defaulting sample, but significantly lower than those of non-defaulting sample. At the same time, the average amounts of income tax and VAT payment in the past six months of non-passing enterprises are significantly higher than those of non-defaulting enterprises. The income tax amount in the first quarter and the second quarter before the loan application, and the VAT amount of the first month, the second month, the third month, and the sixth month before the loan application are also significantly higher than those of the non-defaulting enterprises.

From the above analysis, it can be seen that the amount of loan application, the number of employees, the amount of VAT and income tax paid in the past six months of non-passing enterprise are more similar to those of the non-defaulting enterprises. Some features of non-

passing enterprises, such as the duration, the number of VAT and income tax payments in the past six months, and the number of tax violations, are worse than the relevant features of non-defaulting enterprises, but some features are still better than those of defaulting enterprises. Therefore, the relevant features that can significantly distinguish between defaulting and non-defaulting enterprises cannot reach a consistent conclusion when used to evaluate non-passing enterprises, which indicates that a large part of the non-passing enterprises may be falsely rejected. This thesis intends to establish a more reasonable loan review model to identify them.

Table 4.9 Comparison of the tax features of the three types of samples

Features	Non-defaulting	Defaulting	Non-passing	Non-defaulting – non-passing	P value
Number of tax violations in the past six months	0.159	0.407	0.314	-0.154	0.020**
Number of tax violations in the past year	0.319	0.556	0.612	-0.293	0.011**
Number of tax violations in the past two year	0.571	0.907	1.185	-0.613	0.000***
Number of tax violations in the past three year	0.841	1.185	1.543	-0.702	0.000***
Income tax amount in the first quarter before the loan application	8240	695	9663	-1423	0.000***
Income tax amount in the second quarter before the loan application	7087	1199	9198	-2112	0.000***
VAT amount in the first month before the loan application	22151	5406	22782	-631	0.000***
VAT amount in the second month before the loan application	22838	5565	23715	-877	0.000***
VAT amount in the third month before the loan application	22446	6337	22993	-546	0.000***
VAT amount in the fourth month before the loan application	22285	3941	21410	875	0.000***
VAT amount in the fifth month before the loan application	20153	7121	18990	1163	0.000***
VAT amount in the sixth month before the loan application	19662	5973	20064	-403	0.000***
Number of income tax payments in the past six months	1.33	0.33	0.87	0.46	0.000***
Number of VAT payments in the past six months	4.75	0.94	3.19	1.56	0.000***
Total amount of income tax in the past six months	15326	1894	18861	-3535	0.000***

Total amount of VAT in the past six months	129535	34343	129954	-419	0.000***
--	--------	-------	--------	------	----------

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

4.3 Chapter summary

This chapter introduces the sample data collected in this thesis, and generates new features based on original data, forming a feature set including basic features, financial features and tax features. Then, based on the feature data, this thesis profiles the defaulting sample, initially explores the features that can effectively classify the defaulting and non-defaulting enterprises; and analyzes the characteristics of the non-passing enterprises, and the following findings are obtained.

First, the defaulting enterprises have the following characteristics: in terms of basic features, the loan application amount is lower, and duration is shorter; in terms of financial features, they have higher accounts receivable, prepayments, other receivables and inventories, construction in progress, intangible assets, long-term prepaid expenses, gearing ratio, various liabilities, financial expenses, non-operating income and profits; in terms of tax features, the amount and frequency of their income tax and VAT payments are significantly lower, and enterprises with lower tax credit rating and small-scale taxpayers have a higher probability of default; moreover, their financial information may be unreliable.

Second, some of the features of non-passing enterprises are better than those of non-defaulting enterprises, such as the amount of loan application, the number of employees, the amount of VAT and income tax paid in the past six months. Some features of non-passing enterprises are worse than the relevant features of non-defaulting enterprises, but better than those of defaulting enterprises, such as the duration, the number of VAT and income tax payments in the past six months. It shows that the probability of non-passing enterprises to be falsely rejected is high.

In the next chapter, this thesis will continue to explore the predictive power of the above index system, verify the extent to which the tax features can predict enterprise's default behavior, and use machine learning algorithms to optimize the credit risk evaluation index system of "Tax E Financing".

[This page is deliberately left blank.]

Chapter 5: Optimization of the Credit Evaluation Index System of “Tax E Financing”

This chapter has two research goals. The first one is to verify the predictive power of tax features. For that matter, we will divide the feature data collected in Chapter 4 into multiple feature subsets, use support Vector Machine algorithm to train the classifier for each feature subset, and classify the classification ability of tax features by comparing the performance differences of the classifiers constructed by different datasets. The second one is to further screen and optimize the credit evaluation index system by using the Random Forest algorithm based on an effective and reliable feature set.

We will use Python software and third-party libraries such as “sklearn”, “numpy”, pandas, matplotlib to complete the analysis of this chapter.

5.1 Analysis of the predictive ability of tax features

The tax features are new features proposed in this thesis based on the existing researches, which are used to predict the credit default risk of small and medium-sized enterprises. In order to verify the predictive power of the indicator system, in this thesis we divide the total feature dataset into the following sub-samples. Sample 1 only includes basic features, sample 2 only includes financial features, sample 3 only contains tax features, sample 4 contains basic features and financial features, sample 5 contains basic features and tax features, sample 6 contains financial features and tax features, and sample 7 contains all features, as shown in Table A11. Through the comparative analysis of the predictive ability of these different sub-samples, we are going to assess the validity and reliability of various characteristics in predicting corporate default.

5.1.1 Feature engineering

According to the data processing process of machine learning, in this thesis we need to preprocess each sample first, including outlier processing, missing value processing, coding of categorical features, standardization processing of continuous features, as well as the preliminary screening of indicators.

(1) Outlier processing

By observing the statistical characteristics of each feature, we find that some features do have outliers. For example, some asset items and liability items in the financial characteristics, such as other receivables, inventories, short-term loans, have negative numbers, which is impossible in real data. There are also some features with extreme values, such as the maximum cash ratio of 85,791 times and the maximum inventory turnover rate of 95,032 times.

Considering that the sample size of this thesis is not large, and the number of defaulting enterprises is particularly small, we do not use the 3sigma method to delete samples with outliers, but adopt the winsorize method, that is, samples below the 2.5% quantile and above the 97.5% quantile are replaced by the value of the corresponding quantile instead of deleting them, to ensure that the total number of samples is not reduced.

(2) Missing value processing

By observing the statistical characteristics of each feature, it can be found that there are no missing values in the tax features. In the basic features, only the number of employees and registered capital have missing values of 0.33%, but there are many missing values in the financial features. A total of 56 financial features have missing values, of which the missing rate of inventory is 13.5%, ranking first, which further indicates that the financial information of SMEs is extremely imperfect. The above features with missing values are all continuous variables, and we have chosen to fill in the average value of the feature in each sample.

(3) Coding of categorical features

In the sample data of this thesis, there are five classification features, which are the industry and registration type in the basic features and the tax status name, tax credit rating and tax type in the tax features. In this thesis, one-hot encoding is used to create a dummy variable for each category and convert them into numerical variables.

(4) Standardization processing of continuous features

The standardization of continuous features, also known as dimensionless processing, refers to converting data of different sizes to a uniform size, or converting data of different distributions to a specific distribution. In algorithms centered on gradients and matrices, such as LR, SVM, and BPNN, dimensionless can speed up the solution. while in distance-based models, such as K-nearest neighbor algorithm and K-Means clustering, dimensionless can improve the accuracy of the model and avoid the influence of a feature with a particularly large value range on the distance calculation.

The dimensionless processing of data includes two types: linear processing, and nonlinear processing. In this thesis we use the Z-score normalization method to standardize the data, that

is, after centering the data by the mean, it is then scaled by the standard deviation. The formula is as follows.

$$x^* = \frac{x - \mu}{\sigma} \quad (5.1)$$

(5) Eliminating the invalid indicators

In this thesis, the variance filtering method is used to remove features with a variance of 0 from the data set. Because a variance of 0 means that the feature has no difference between defaulting and non-defaulting enterprises, therefore it cannot play any role in the classification.

5.1.2 Data visualization

After completing the data preprocessing of each sub-sample, we will use Principal Component Analysis (PCA) to reduce the dimensionality of the features of each data set, and extract the first two or three components. By observing the distribution of defaulting and non-defaulting enterprises in two-dimensional and three-dimensional space, we are able to preliminarily judge whether the data set has linear separability, that is, whether a line or a plane can be found to effectively separate defaulting enterprises from non-defaulting enterprises.

(1) Data visualization of sample 1

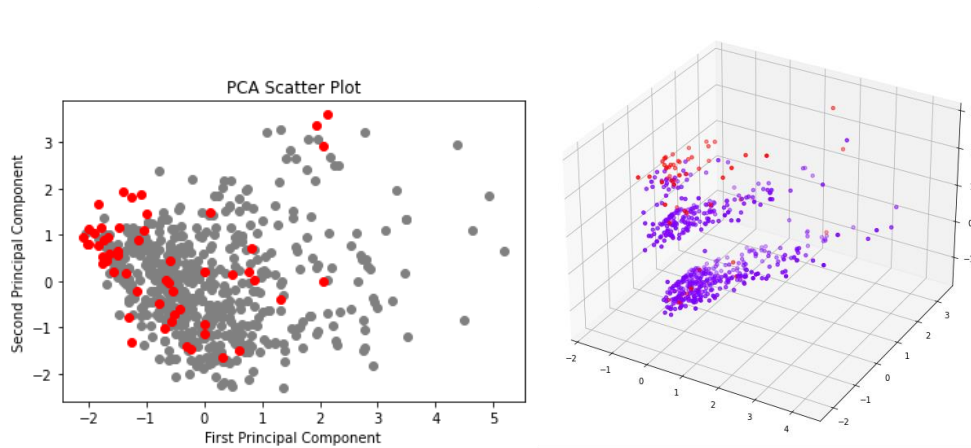


Figure 5.1 2D and 3D visualization of sample 1

Source: Author.

Figure 5.1 shows the discriminability of sample 1 (which only contains basic features) between defaulting and non-defaulting samples in 2D and 3D space. The red dots in the figure represent defaulting enterprises. It can be seen from the figure that when only the first two principal component information is extracted, the distribution of defaulting enterprises is more concentrated than that of non-defaulting enterprises, but the two samples have high overlap and low linear separability. However, in the three-dimensional space composed of the three principal components, the defaulting enterprises obviously have a higher third principal

component value. From that 3D figure, the two types of enterprises have strong linear separability.

(2) Data visualization of sample 2

Figure 5.2 shows the discriminability of sample 2 (which only contains financial features) between defaulting and non-defaulting samples in 2D and 3D space. It can be found from the figure that, either in two-dimensional or three-dimensional space, the degree of mixing of defaulting enterprises and non-defaulting enterprises is high, and the degree of linear differentiation is low.

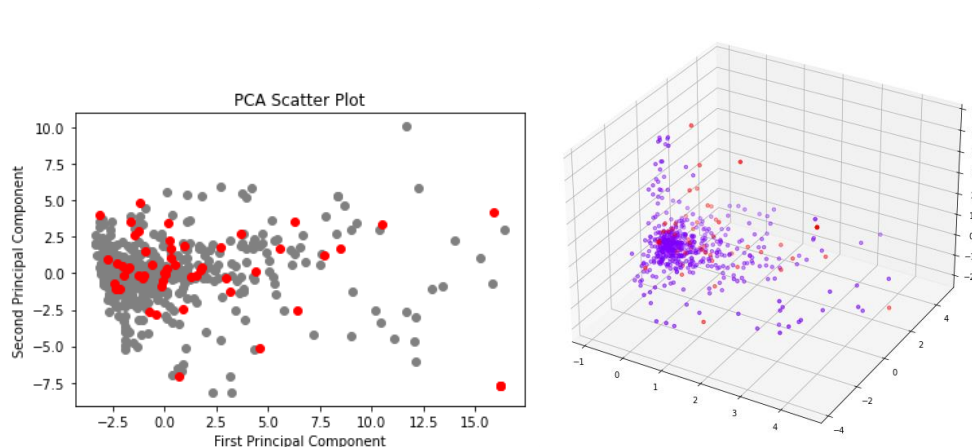


Figure 5.2 2D and 3D visualization of sample 2

Source: Author.

(3) Data visualization of sample 3

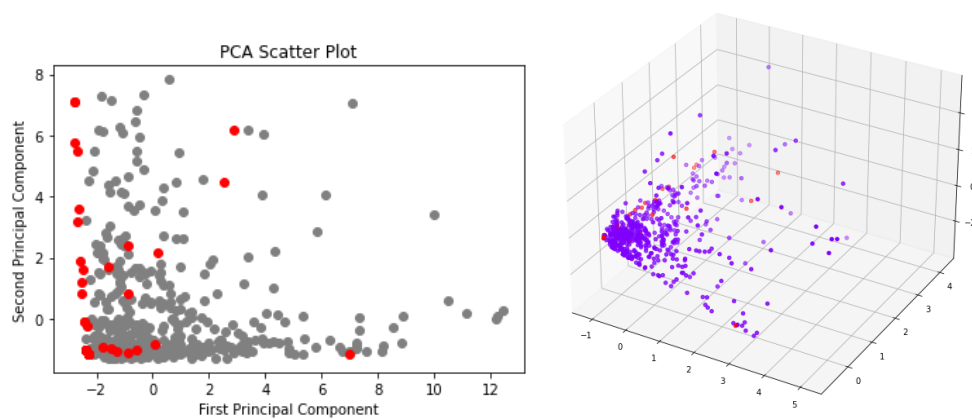


Figure 5.3 2D and 3D visualization of sample 3

Source: Author.

Figure 5.3 shows the discriminability of sample 3 (which only contains tax features) between defaulting and non-defaulting samples in 2D and 3D space. It can be seen from the figure that the distribution of defaulting enterprises is relatively concentrated, their first and second component values are relatively low, and the differences in the third component values are also relatively small. Therefore, the distributions of defaulting enterprises in two-

dimensional space and three-dimensional space are very similar.

(4) Data visualization of sample 4

Figure 5.4 shows the discriminability of sample 4 (which contains basic features and financial features) between defaulting and non-defaulting samples in 2D and 3D space. It can be seen that both in the two-dimensional space and in the three-dimensional space, the defaulting enterprises and the non-defaulting enterprises do not show good linear separability. Compared with Figure 5.1, which only includes basic features, Figure 5.4 shows relatively poor linear separability, which indicates that adding financial features to basic features not only does not significantly improve the linear separability between defaulting enterprises and non-defaulting enterprises, but instead increases data noise, resulting in more linear inseparability.

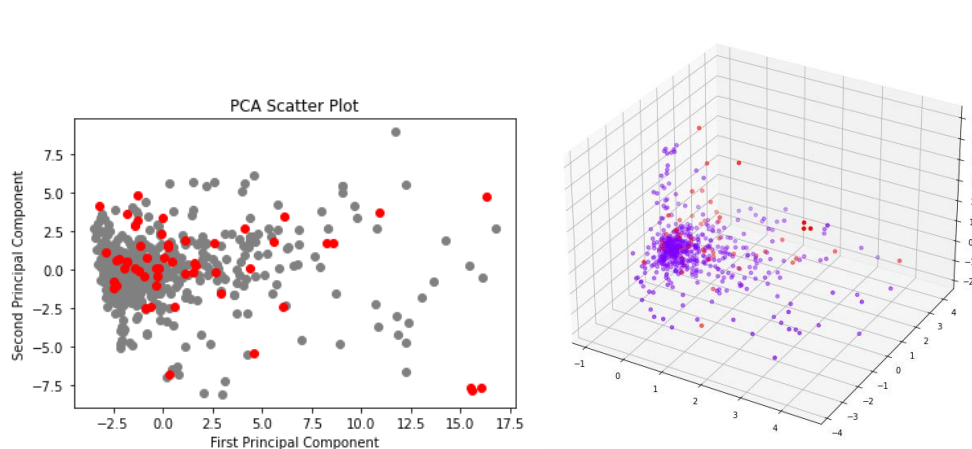


Figure 5.4 2D and 3D visualization of sample 4

Source: Author.

(5) Data visualization of sample 5

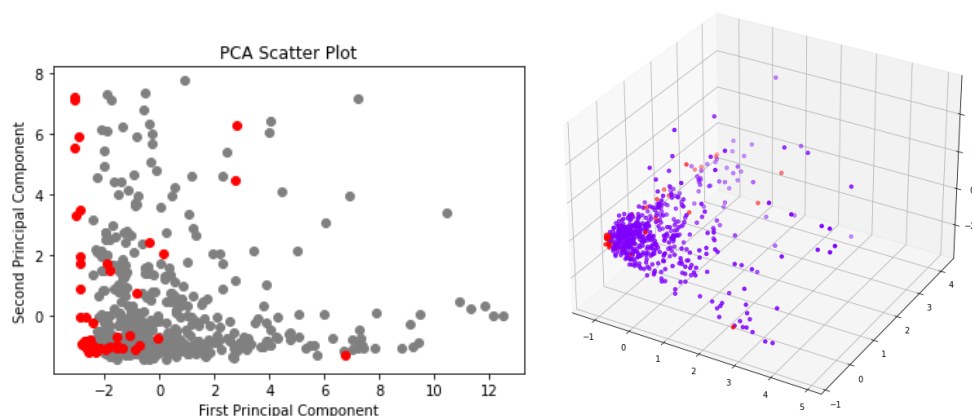


Figure 5.5 2D and 3D visualization of sample 5

Source: Author.

Figure 5.5 shows the discriminability of sample 5 (which contains basic features and tax features) between defaulting and non-defaulting samples in 2D and 3D space. Comparing Figure 5.2 and Figure 5.5, we can find that the 2D graphs of the two are very similar, but by

observing the 3D graphs, it can be found that the combination of tax features and basic features will make the distribution of defaulting enterprises more concentrated and make it easier to use a linear plane to separate them from non-defaulting enterprises.

(6) Data visualization of sample 6

Figure 5.6 shows the discriminability of sample 5 (which contains tax features and financial features) between defaulting and non-defaulting samples in 2D and 3D space. Comparing it with Figure 5.2, we can find that the distribution of defaulting enterprises is more concentrated in both the two-dimensional space and the three-dimensional space. While compared with Figure 5.3, it can be found that whether in two-dimensional space or three-dimensional space, the distribution of defaulting enterprises is more dispersed and the degree of linear separability with non-defaulting enterprises is lower. This suggests that tax features can improve classification ability, while financial features will reduce classification ability.

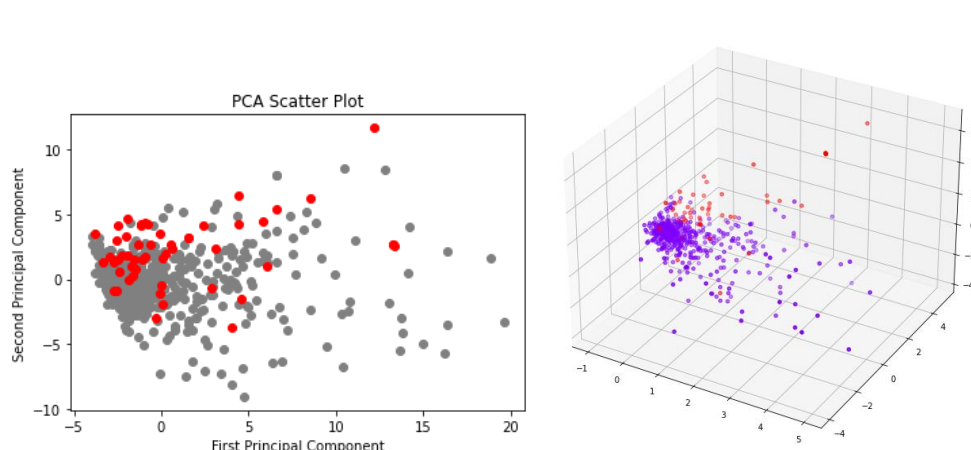


Figure 5.6 2D and 3D visualization of sample 6

Source: Author.

(7) Data visualization of sample 7

Figure 5.7 shows the discriminability of sample 5 (which contains all the features) between defaulting and non-defaulting samples in 2D and 3D space. Comparing Figure 5.7 with Figure 5.6, we can find that adding basic features can generally increase the value of the third principal component of the defaulting enterprises and move their positions in the three-dimensional space upward, which is beneficial to increase the linear separability. Comparing Figure 5.7 with Figure 5.5, we can find that after adding financial features, the distribution of defaulting enterprises is more dispersed, and the degree of linear separability with non-defaulting enterprises decreases.

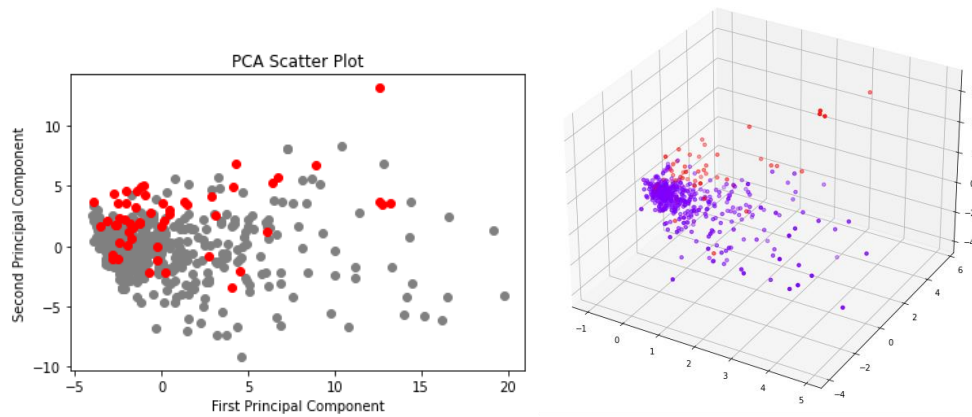


Figure 5.7 2D and 3D visualization of sample 7

Source: Author.

Through the data visualization analysis of 7 sub-samples in 2D and 3D space, the following findings are found.

First, basic features and tax features on their own have good linear separability for defaulting enterprises and non-defaulting enterprises.

Second, the combination of basic features and tax features increases the linear separability of data.

Third, the linear separability of financial features for defaulting and non-defaulting enterprises is low. After mixing it with either basic features or tax features, not only it does not increase, but it actually decreases the degree of linear separability.

This preliminary analysis shows that the financial features are not suitable for the credit risk evaluation of “Tax E Financing”, but this analysis is only a graphical analysis from the perspective of linear separability, and there are many nonlinear classification models that can be trained. Therefore, in the following, we will comprehensively use the linear and nonlinear classification models to test the classification ability of the above 7 data sets, and clarify which kind of features have better predictive ability for the credit risk of “Tax E Financing” enterprises.

5.1.3 Predictive ability comparison of SVM classifiers on 7 sub-samples

In this section, the classifier will be trained on 7 sub-samples with SVM. Firstly, in this thesis we choose Radial Basis Function kernel (RBF) as the kernel function of SVM, and the reason was explained in 3.3.1. Secondly, due to the extremely uneven distribution of data, the ratio of defaulting enterprises and non-defaulting enterprises is 1:9.75. In order to capture more of the minority class, namely defaulting enterprises, in the training process of the classifier, we set the “class_weight” parameter in the SVC class of the sklearn library to “balanced”, that is, input

according to the true ratio of defaulting enterprises and non-defaulting enterprises. If this parameter is not set, it defaults to 1:1, which will reduce the probability of minority class samples being captured.

In addition to the above two points, the follow-up process also includes parameter optimization, sample splitting, model training and result analysis, determination of optimal thresholds and reclassification. In the following, we will show the above processes and analyze the related results step by step.

5.1.3.1 Parameter optimization

As shown in Table A3, the RBF kernel comes with a parameter σ , and the objective function of SVM algorithm also has a penalty coefficient C , so we need to determine the optimal values of these two parameters first. The goal of optimization is to maximize the AUC value of the classifier. In this thesis we will use grid search and 5-fold cross-validation to optimize the parameters C and σ .

Grid search is to combine the possible values of each parameter and list all possible combinations to generate a “grid”. Each combination is then used to train a classifier with SVM, and the performance of the classifier is evaluated using cross-validation. According to the experience, the initial value range of C is $[0.1, 20]$, and the initial value range of σ is $[-10, 1]$. After the first optimization, it will continue to perform finer values near the first optimal parameter, and perform the second optimization until the model AUC value no longer changes. 5-fold cross-validation refers to dividing each data set into 5 parts, taking four of them as the training set and the remaining one as the testing set, performing 5 training and testing, and returning the average AUC value of the 5 learners.

Through the above process, we use SVM to optimize the parameters of 7 sub-samples, and the final selected values of C and σ are shown in Table A12.

5.1.3.2 Sample splitting

After obtaining the optimal parameter values of C and σ , we will randomly split each sample using the hold-out method, and select 70% of the sample as the training set and the remaining 30% of the sample as the testing set. Then use SVM to train the training set data and perform performance testing on the testing set. Since the enterprises of each sub-sample are the same, only the features of each sample are different. In order to enhance the comparability of the prediction results, we set the same random parameters when splitting the samples to ensure the training set and testing set of each sub-sample are consistent. The split results are shown in

Table A13.

It can be seen from the table that both the training set and the testing set contain non-defaulting enterprises and defaulting enterprises, and the proportions of the two are relatively balanced. The testing set contains a total of 180 enterprises, with defaulting enterprises accounting for 10.6%. This ratio can be understood as the actual default rate on the testing set, which will be used to compare with the predicted default rate of the classifier below.

5.1.3.3 Performance comparison

(1) Classification result of sample 1

Table A14 shows the confusion matrix of the SVM classifier based on the basic features, and Table A15 shows the various performance indicators of the classifier on the testing set. In addition to the commonly used indicators such as Accuracy, Recall and Precision, in this thesis we also calculate the approval rate and default rate to evaluate the performance of the classifier. Among them, the approval rate is calculated using $(FN+TN)/(FN+TN+TP+FP)$, which reflects the proportion of the enterprises predicted to be “non-defaulting” to the total number of enterprises in the testing set, and the default rate is calculated using $FN/(FN+TN)$, reflecting the proportion of enterprises that are predicted to be “non-defaulting” but actually “defaulting” to the total number of enterprises that are predicted to be “non-defaulting”. For the classifier of “Tax E Financing” constructed in this thesis, we hope to achieve a balance between the “approval rate” and the “default rate”, that is, the balance between the recall rate and the precision rate. We should not only find out defaulting enterprises, but also not mistakenly reject too many non-defaulting ones.

It can be found from these tables that the basic features in sample 1 have extremely high predictive ability for defaulting enterprises. All 19 defaulting enterprises in the testing set are identified, and the recall rate reaches 100%. However, among the 161 non-defaulting enterprises, only 98 are classified correctly and 63 enterprises are wrongly classified as defaulting. Therefore, the accuracy rate and precision rate of the classifier are low, with the accuracy rate of only 65% and the precision rate of only 23.17%. If this classifier is used to review loans on the testing set, the approval rate is only 54.44%, and the default rate is 0%. It should be noted that the enterprises in the testing set have passed the loan approval under the “one-vote veto system” loan review mechanism of “Tax E Financing”. When considering the non-passing enterprises that are rejected by the rules, the actual approval rate will drop significantly. Therefore, we can find that the SVM classifier based on basic features does not achieve a good balance between the accuracy rate and the precision rate, and also the approval

rate and the default rate.

Figure B5 shows the ROC curve of the SVM classifier based on sample 1. The horizontal axis of the curve shows the false positive rate (FPR), that is, the rate at which enterprises that are non-defaulting are considered to be defaulting, and the vertical axis shows the recall rate, that is, the proportion of enterprises that are actually defaulting being identified. The curve takes the distance between each sample point and the hyperplane as the sorting variable, sorts them in descending order, and uses each value as a threshold in turn. The samples whose distance is greater than or equal to the threshold are classified as negative, that is, they will not default, and the rest of the samples are classified as positive, which means they will default, and calculate the FPR and Recall under each threshold in turn, and connect all the points to get the ROC curve. Therefore, each point on the ROC curve corresponds to a threshold. It can be found from Figure B5 that when the FPR is around 0, a certain threshold division of the SVM classifier can achieve a recall rate of about 80%. We can also find the point closest to the upper left corner on the entire ROC curve, or the point farthest from the black dotted line, set the threshold corresponding to this point as the optimal threshold, and reclassify the samples according to this threshold. At this time, we will abandon the classification according to the symbol in the original SVM model, that is, the classification standard shown in formula (3.19), and instead classify the samples according to the optimal threshold.

Through the search, the optimal threshold for Figure B5 is 0.7686, and the results after reclassifying the testing set according to this threshold are as follows. Table A16 shows the confusion matrix for classification, and Table A17 shows the related performance metrics. It can be found that after using the optimal threshold classification, the recall rate of the model decreases from 100% to 89.47%, but the accuracy rate of the model increases from 65% to 92.22%. 149 out of 161 non-defaulting enterprises are accurately identified, and 17 out of 19 defaulting enterprises are accurately identified. The approval rate of the model rises from 54.44% to 83.89%, and the default rate rises from 0 to 1.32%, but it is still at a low level. The two achieve a good balance.

(2) Classification result of sample

Table A18 shows the confusion matrix of the SVM classifier based on financial features, and Table A19 shows various performance indicators of the classifier on the testing set. As can be seen from these tables, the overall accuracy rate of the classifier trained with financial features is 91.11%, which is higher than the accuracy rate of sample 1. However, the classifier's ability to identify defaulting enterprises is relatively poor. Only 13 of the 19 defaulting enterprises are identified, and the recall rate is only 68.42%, which is only slightly higher than

the 50% recall rate of random identification. The approval rate of 87.22% is also higher than in sample 1, but the default rate has also risen sharply to 3.82%.

Figure B6 shows the ROC curve of the SVM classifier based on sample 2. It can be seen that the AUC area under the ROC curve is only 86.56%, which is less than 96.5% of the AUC area under Figure B5. Similar to the previous content, we can also find the point closest to the upper left corner on the entire ROC curve to determine the optimal threshold, and reclassify the samples according to this threshold. After calculation, we can find the optimal threshold is -0.2417.

Table A20 shows the confusion matrix after classification according to the optimal threshold, and Table A21 shows the related performance metrics. It can be found that after using the optimal threshold classification, the recall rate of the model increases from 68.42% to 78.95%, but the accuracy rate of the model decreases from 91.11% to 88.89%. 145 out of 161 non-defaulting enterprises are accurately identified, and 15 out of 19 defaulting enterprises are accurately identified. The approval rate for the model falls from 87.22% to 82.78%, and the default rate falls from 3.82% to 2.68%. After using the optimal threshold classification, the balance between recall rate and accuracy rate becomes better. However, it can be found that the optimal accuracy and recall rate that can be achieved by using financial features are worse than the results of classification using basic features, the approval rate is lower, and the default rate is higher.

(3) Classification result of sample 3

Table A22 shows the confusion matrix of the SVM classifier based on the tax features, and Table A23 shows the various performance indicators of the classifier on the testing set. As can be seen from the table, 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%, and 158 of the 161 non-defaulting enterprises can be correctly identified. The overall accuracy rate of the classifier trained with tax features is 97.22%, which is much higher than the accuracy of the SVM classifier trained with basic features and financial features. Before using the optimal threshold classification, the approval rate of the model reaches 88.89%, and the default rate is 1.25%, which is the best performance that the first three samples can achieve.

Figure B7 shows the ROC curve of the SVM classifier based on sample 3. It can be seen that the AUC area under the ROC curve is 97.03%, which is higher than the AUC area in Figure B6 and Figure B5, which indicates that the tax features distinguish the defaulting enterprises from non-defaulting enterprises better than basic features and financial features. We still look for the point closest to the upper left corner on the entire ROC curve, and the optimal threshold

is 0.9996.

We reclassify the testing set by this threshold. Table A24 shows the confusion matrix after classification according to the optimal threshold, and Table A25 shows the related performance metrics. It can be found that after using the optimal threshold for classification, the recall rate of the model remains unchanged at 89.47%, but the accuracy rate of the model increases from 97.22% to 97.78%. 159 out of 161 non-defaulting enterprises are accurately identified, and 17 out of 19 defaulting enterprises are accurately identified. The approval rate for the model rises from 88.89% to 89.44%, and the default rate falls from 1.25% to 1.24%. After using the optimal threshold for classification, the approval rate and default rate of the model both become better.

Comparing the performance indicators of the SVM classifiers of sample 1, sample 2 and sample 3, it can be seen that the SVM classifier trained with tax features has the highest approval rate and the lowest default rate, followed by basic features, and the worst is financial features. The above analysis effectively verifies the ability of tax features to classify defaulting enterprises and non-defaulting enterprises. In order to test whether the combination of different types of features can improve the classification ability of the model, in this thesis we continue to train and test the SVM classifier on the remaining samples.

(4) Classification result of sample 4

Table A26 shows the confusion matrix of the SVM classifier based on basic features and financial features, and Table A27 shows various performance indicators of the classifier on the testing set. It can be seen from the table that 13 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 68.42%, which is at a low level. Among the 161 non-defaulting enterprises, 156 can be correctly identified. The overall accuracy rate of the classifier trained with sample 4 is 93.89%, which is lower than the accuracy rate of the SVM classifier trained with sample 3. The approval rate of the model reaches 90%, which is the highest value among the first four samples, but the default rate is also at a high level, reaching 3.7%, which is only slightly lower than the default rate of the SVM classifier trained with financial features.

Figure B8 shows the ROC curve of the SVM classifier based on sample 4. It can be seen that the AUC area under the ROC curve is 93.04%, which is higher than the AUC of sample 2, but lower than the AUC of samples 1 and 3. It can be seen that after adding financial features to the basic features, its ability to classify defaulting enterprises and non-defaulting enterprises is better than using only financial features, but it is worse than using only basic features. We still look for the point closest to the upper left corner on the entire ROC curve to determine the optimal threshold, which is -0.0355.

We reclassify the testing set according to the optimal threshold. Table A28 shows the

confusion matrix after classification according to this threshold, and Table A29 shows the related performance metrics. It can be found that after using the optimal threshold for classification, the recall rate of the model increases from 68.42% to 78.95%, and the accuracy rate also increases from 93.89% to 94.44%. 155 out of 161 non-defaulting enterprises are accurately identified, and 15 out of 19 defaulting enterprises are accurately identified. The approval rate for the model falls from 90% to 88.33%, and the default rate falls from 3.7% to 2.52%. It can be found that after using the optimal threshold for classification, the identification rate of sample 4 is significantly lower than that of sample 1 and sample 3, and the accuracy rate is also lower than that of sample 3.

(5) Classification result of sample 5

Table A30 shows the confusion matrix of the SVM classifier based on basic features and tax features, and Table A31 shows various performance indicators of the classifier on the testing set. It can be seen from the table that 15 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 78.59%, and 160 of the 161 non-defaulting enterprises can be correctly identified. The overall accuracy of the classifier trained with sample 5 is 97.22%, which is comparable to that of sample 4, but the recall is slightly lower than that of sample 4. The approval rate of the model reaches 91.11%, the highest value among the first five samples, but the default rate is also at a high level, reaching 2.44%, slightly higher than the default rate of sample 4.

Figure B9 shows the ROC curve of the SVM classifier based on sample 5. It can be seen that the AUC area under the ROC curve is 94.21%, which is slightly lower than the AUC of sample 4. It can be seen that under the original classification standard of SVM, adding basic features to the tax features reduces the model's ability to classify defaulting enterprises and non-defaulting enterprises, especially the ability to identify defaulting enterprises. We still look for the point closest to the upper left corner on the entire ROC curve, and the threshold corresponding to this point is calculated to be -0.2575.

In this thesis we reclassify the testing set according to this threshold. Table A32 shows the confusion matrix after classification according to the optimal threshold, and Table A33 shows the related performance metrics. It can be found that after using the optimal threshold for classification, the recall rate of the model increases from 78.95% to 89.47%, and the accuracy rate also increases from 97.22% to 97.78%. 159 out of 161 non-defaulting enterprises are accurately identified, and 17 out of 19 defaulting enterprises are accurately identified. The approval rate for the model falls from 91.11% to 89.44%, and the default rate falls from 2.44% to 1.24%. It can be found that after using the optimal threshold for classification, the

performance indicators of sample 5 are exactly the same as those of sample 3, and various indicators are higher than those of sample 1, sample 2 and sample 4.

(6) Classification result of sample 6

Table A34 shows the confusion matrix of the SVM classifier based on tax features and financial features, and Table A35 shows various performance indicators of the classifier on the testing set. It can be seen from the table that 16 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 84.21%, and 158 of the 161 non-defaulting enterprise can be correctly identified. The overall accuracy of the classifier trained with sample 6 is 96.67%. Its accuracy rate and recall rate both are higher than sample 2, but lower than sample 4. The approval rate of the model reaches 89.44%, and the default rate is 1.86%. At this time, compared with sample 3, the approval rate is slightly higher, but the default rate is also higher.

Figure B10 shows the ROC curve of the SVM classifier based on sample 6. It can be seen that the area under the ROC curve is 93.59%, which is lower than that of sample 4. It can be seen that adding financial features to tax features actually reduces the data's ability to classify defaulting enterprises and non-defaulting enterprises. We still look for the point closest to the upper left corner on the entire ROC curve, and the threshold corresponding to this point is -0.0014.

We reclassify the testing set according to this threshold. Table A36 shows the confusion matrix after classification according to the optimal threshold, and Table A37 shows the related performance metrics. It can be found that after using the optimal threshold for classification, the recall rate of the model increases from 84.21% to 89.47%, and the accuracy rate also increases from 96.67% to 97.22%. 158 out of 161 non-defaulting enterprises are accurately identified, and 17 out of 19 defaulting enterprises are accurately identified. The approval rate for the model falls from 89.44% to 88.89%, and the default rate falls from 1.86 %to 1.25%. We can find that after using the optimal threshold for classification, the performance of sample 6 is still worse than that of sample 3, the default rate is higher, the approval rate is lower, and the recall rate is the same, but the accuracy rate is lower.

(7) Classification result of sample 7

Table A38 shows the confusion matrix of the SVM classifier based on basic features, tax features and financial features, and Table A39 shows various performance indicators of the classifier on the testing set. It can be seen from the table that 12 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 63.16%, and 160 of the 161 non-defaulting enterprises can be correctly identified. The overall accuracy rate of the classifier trained with sample 7 is 95.56%, the approval rate is 92.78%, and the default rate also reaches 4.19%, both

of which reach the highest value among the seven models.

Figure B11 show the ROC curve of the SVM classifier based on sample 7. It can be seen that the area under the ROC curve is 95.85%, which is still lower than that of sample 4. We still look for the point closest to the upper left corner on the entire ROC curve, and the threshold corresponding to this point is found to be -0.5655.

In this thesis we reclassify the testing set according to this threshold. Table A40 shows the confusion matrix after classification according to the optimal threshold, and Table A41 shows the relevant performance indicators. It can be found that after using the optimal threshold for classification, the recall rate of the model increases from 63.16% to 89.47%, and the accuracy rate decreases from 95.56% to 90%. 145 out of 161 non-defaulting enterprises are accurately identified, and 17 out of 19 defaulting enterprises are accurately identified. The approval rate for the model falls from 92.78% to 81.67%, and the default rate falls from 4.19% to 1.36%. It can be found that after using the optimal threshold classification, the model's identification accuracy for defaulting enterprises is improved, but the ability to identify non-defaulting enterprises is reduced. The recall rate of sample 7 is the same as that of samples 1, 3, 5, and 6, but the accuracy rate is significantly lower than them, and while the approval rate of sample 7 is lower than them, on the contrary the default rate is higher than them. This shows that mixing all the features together does not increase the classification ability of the model, but the performance index decreases significantly.

5.1.4 Summary

Figure 5.8 shows the main performance indicators of 7 sub-samples after classification using the optimal threshold, including accuracy rate, recall rate, approval rate and default rate.

By comprehensively comparing the relevant indicators of the 7 sub-samples, the following conclusions can be drawn.

First, sample 3 (tax features) and sample 5 (basic features + tax features) have the best classification performance, and their performance is the same. Among all feature subsets, sample 3 and sample 5 have the highest accuracy rate, recall rate, and approval rate, and the lowest default rate. Specifically, they have a classification accuracy rate of 97.78% for defaulting enterprises and non-defaulting enterprises, and a recall rate of 89.47% for defaulting enterprises. According to this classifier, the loan approval rate of the enterprises in the testing set is 89.44%, and the default rate is 1.24%.

Second, the performance index of sample 2 (financial features) is the worst, and the SVM

classifier trained based on this dataset has the lowest accuracy rate and recall rate, and the highest default rate. After mixing financial features with basic features and tax features, and then training the SVM classifier, the performance of the model is better than only using financial features, but worse than not using financial features. This shows that adding financial features to the “Tax E Financing” credit evaluation index system will not only fail to play a positive role in the distinction between defaulting enterprises and non-defaulting enterprises, but will rather have a negative impact. This further verifies the findings of this thesis in Section 5.1.2, and also proves the unreliability of the financial features of SMEs.

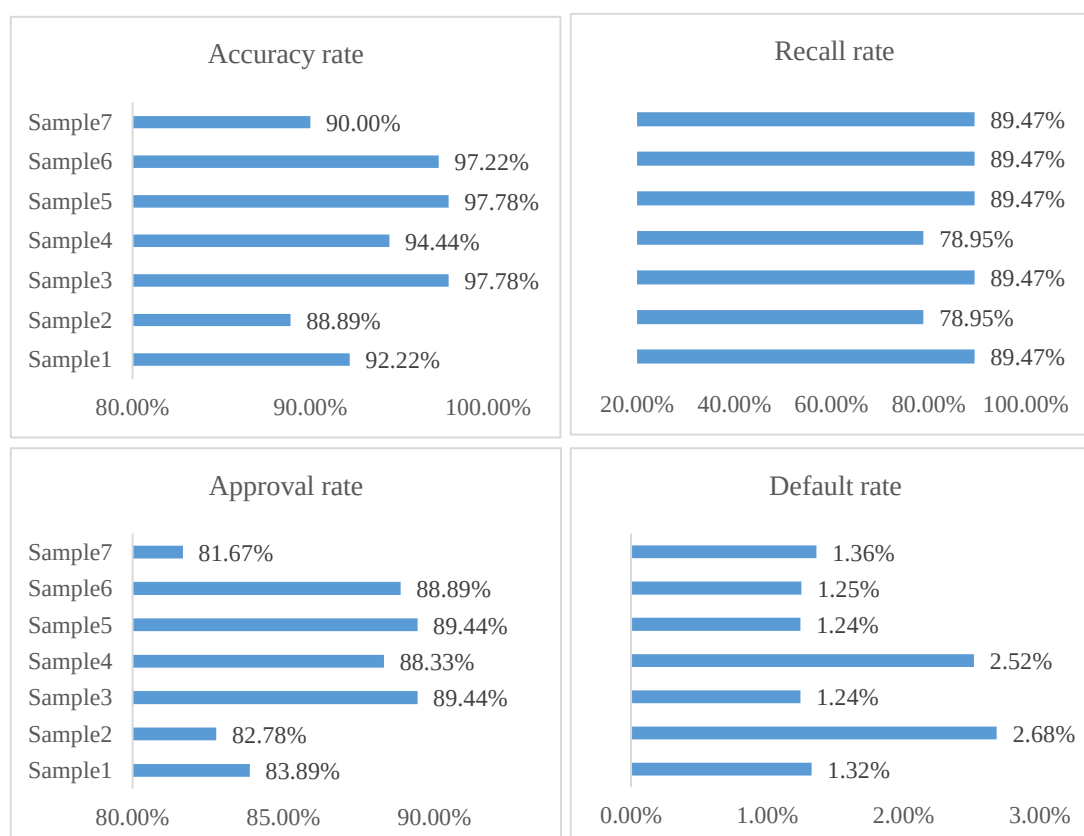


Figure 5.8 Performance comparison of SVM classifier on 7 sub-samples

Source: Author.

5.2 Credit evaluation index system optimization based on reliable information

The analysis in Section 5.1 finds that financial information is provided by small and medium-sized enterprises themselves, which is incomplete and unreliable, and has poor performance in credit risk prediction. The basic information and tax information are collected and recorded by the government’s industrial and commercial departments and tax departments, which are authentic and reliable. The analysis above also proves that they do have a high predictive ability

for the credit risk of “Tax E Financing” enterprises. Since the above analysis found that samples 3 and 5 have the best and the same performance indicators, in the following, we will use the Random Forest algorithm to further optimize the features contained in samples 3 and 5, select the features that have important contributions to the performance indicators, and delete the irrelevant features. The specific process is as follows.

First, use the random Forest algorithm to train sample 3 and sample 5, and return the feature importance of the model. The calculation process of this indicator is shown in 3.3.2.2. Sort the feature importance in descending order, take each value as the threshold, delete the features smaller than the threshold, use the remaining features to form a data set, and return the average accuracy of 5-fold cross-validation.

Second, draw the learning curve of threshold and accuracy, select the threshold corresponding to the maximum average accuracy as the optimal threshold, delete features whose importance is less than the threshold, and form the remaining features into a new dataset.

5.2.1 Optimization of tax features

We firstly optimize sample 3, that is, the tax features. the Random Forest algorithm is used to train the data. Among them, in order to improve the efficiency of model operation, the parameter `n_estimators` of `RandomForestClassifier` in `sklearn` is set to 10, which means that 10 decision trees are generated, and the Gini index is used to select branch attributes. The rest of the parameters are default values.

(1) Returns the feature importance values

Sample 3 contains a total of 22 tax features, and the importance of each feature returned by the Random Forest algorithm is shown in Table 5.1. It can be seen from the table that the index that contributes the most to the accuracy of the model is “Number of VAT payments in the past six months” with an importance of 24.55%, followed by “VAT amount in the first month before the loan application”. There are 9 indicators whose importance exceeds 5%, and the importance of “Tax status” is 0, which means that it has no influence on the classification effect. The effect of “Taxpayer type” is also very small.

(2) Find the optimal threshold

Further, we plot the learning curve of the importance threshold and the average accuracy rate of cross-validation, as shown in Figure B12. It can be seen that, as the threshold increases, more and more features are deleted, and the accuracy of the model shows a downward trend. When the threshold exceeds 0.1, the accuracy of the model reaches the lowest, and when the

threshold is around 0.01, the accuracy of the model reaches the highest. By narrowing the scope and redrawing the learning curve, we can find that the optimal threshold is 0.01416. At this time, we will delete the features whose importance is less than the threshold, that is, the “Number of tax violations in the past six months” and the following features in Table 5.1 will be deleted. The number of features has dropped from 22 to 13, which is called new sample 3 in the following.

Table 5.1 Feature importance of sample 3

Feature	Importance
Number of VAT payments in the past six months	24.550%
VAT amount in the first month before the loan application	9.739%
Total amount of income tax in the past six months	9.175%
VAT amount in the second month before the loan application	8.307%
Total amount of VAT in the past six months	8.251%
Income tax amount in the second quarter before the loan application	6.511%
Number of income tax payments in the past six months	6.246%
VAT amount in the third month before the loan application	6.038%
VAT amount in the fifth month before the loan application	5.296%
VAT amount in the sixth month before the loan application	4.786%
Number of tax violations in the past three year	3.551%
Number of tax violations in the past two year	1.698%
Number of tax violations in the past year	1.416%
Number of tax violations in the past six months	1.387%
Income tax amount in the first quarter before the loan application	0.959%
Tax credit rating =A	0.951%
VAT amount in the fourth month before the loan application	0.844%
Tax credit rating =B	0.281%
Taxpayer type = General taxpayer	0.008%
Taxpayer type = Small-scale taxpayer	0.007%
Tax status = normal	0.000%
Tax status = not normal	0.000%

Source: Author.

5.2.2 Optimization of basic features and tax features

We then optimize sample 5, which is a collection of basic features and taxation features.

(1) Returns the feature importance values

Sample 5 contains a total of 32 features, and the importance of each feature returned by the Random Forest algorithm is shown in Table 5.2. It can be seen from the table that the index that contributes the most to the accuracy of the model is “Total amount of VAT in the past six months”, with an importance of 20.786%, followed by “Loan application amount”, with an importance of 15.522%. The importance of “Tax status” and “Taxpayer type” is 0, which means that they have no influence on the classification effect.

(2) Find the optimal threshold

In this thesis we plot the importance threshold versus the average cross-validation accuracy, as shown in Figure B13. The calculation shows that the optimal threshold corresponding to the highest average accuracy rate is 0.00724. At this point, we will delete features whose importance is less than this threshold, i.e. “Industry = Manufacturing” and the following features in Table 5.2. The number of features of the new sample is reduced from the previous 32 to 20, which is called new sample 5 in the following of this thesis. The 20 features contain 5 basic features and 15 tax features, of which 12 tax features are also included in the new sample 3.

Table 5.2 Feature importance of sample 5

Feature	Importance
Total amount of VAT in the past six months	20.786%
Loan application amount	15.522%
Number of VAT payments in the past six months	7.034%
VAT amount in the third month before the loan application	5.791%
VAT amount in the sixth month before the loan application	5.319%
VAT amount in the second month before the loan application	5.149%
VAT amount in the fourth month before the loan application	4.813%
Income tax amount in the first quarter before the loan application	4.602%
Number of income tax payments in the past six months	4.241%
Enterprise's duration	3.644%
Total amount of income tax in the past six months	3.472%
VAT amount in the first month before the loan application	2.807%
Number of tax violations in the past three year	2.607%
VAT amount in the fifth month before the loan application	2.518%
Registered capital	2.420%
Number of employees	2.093%
Income tax amount in the second quarter before the loan application	1.713%
Number of tax violations in the past two year	1.064%
Registration type = Private limited company	0.751%
Number of tax violations in the past six months	0.724%
Industry = Manufacturing	0.703%
Tax credit rating = A	0.574%
Tax credit rating = B	0.574%
Registration type = Other limited company	0.409%
Number of tax violations in the past year	0.274%
Industry = Wholesale and retail	0.254%
Industry = Other industries	0.193%
Tax status=normal	0.000%
Tax status =not normal	0.000%
Taxpayer type = General taxpayer	0.000%
Taxpayer type = Small-scale taxpayer	0.000%
Registration type = Sole proprietorship	0.000%

Source: Author.

5.3 Chapter summary

This chapter divides the dataset collected in Chapter 4 into multiple sub-samples. For each sub-sample, SVM is used to train the classifier. By comparing the performance differences of the classifiers trained with different datasets, the classification ability of tax features is clarified. Then, we select feature subsets with more classification power from the seven feature subsets, and use the Random Forest algorithm to further screen and optimize the index system. The following findings were made.

First, in terms of the classification performance of each sub-sample, Sample 3 (tax features) and sample 5 (basic features and tax features) have the best classification performance, and the accuracy rate, recall rate, and approval rate of both have reached the highest and their default rates are the lowest. The performance of sample 2 (financial features) is the worst. After mixing financial features with basic features and tax features, not only can it not play a positive role in the distinction between defaulting enterprises and non-defaulting enterprises, but it will have a negative impact.

Second, in terms of optimization and screening of indicators, since sample 3 and sample 5 have the same classification performance, in this thesis we use Random Forest algorithm to optimize these two feature sets at the same time, and eliminate features whose importance is lower than the threshold. After feature optimization of sample 3, 13 more important tax features are extracted and named as new sample 3, and after feature optimization of sample 5, 20 more important basic features and tax features are extracted and named for the new sample 5. 12 of the 13 tax features of the new sample 3 are included in the new sample 5.

In the next chapter, we will use various algorithms to train and test the new sample 3 and the new sample 5, trying to select the optimal classification algorithm for “Tax E Financing”.

Chapter 6: Credit evaluation Method Selection of “Tax E Financing”

Based on the new sample 3 (tax features) and new sample 5 (basic features and tax features), this chapter will use various supervised learning algorithms and semi-supervised learning algorithms to train and test the classifier of “Tax E Financing”. Among them, the supervised learning algorithms include individual algorithms and ensemble algorithms. Individual algorithms such as SVM, LR and BPNN, ensemble algorithms such as RF, and semi-supervised learning algorithms such as TSVM will be used in this thesis. These algorithms are widely used in existing research. We will compare the performance of the classifiers trained by each algorithm, and select the classifier with the best performance as the credit evaluation model of “Tax E Financing”. Meanwhile, in order to increase the comparability of the performance of different algorithms, when splitting the training set and the testing set, the random splitting method consistent with Chapter 5 is adopted, so that this thesis can compare the performance of the SVM classifier before and after the index optimization, and compare the performance of classifiers trained by different algorithms based on the optimized index system.

Indicators contained in new sample 3 and new sample 5 are shown in Table 6.1.

Table 6.1 The optimized credit evaluation index system

New sample 3(tax features)	New sample 5(tax features+ basic features)
Number of VAT payments in the past six months	Total amount of VAT in the past six months
VAT amount in the first month before the loan application	Loan application amount
Total amount of income tax in the past six months	Number of VAT payments in the past six months
VAT amount in the second month before the loan application	VAT amount in the third month before the loan application
Total amount of VAT in the past six months	VAT amount in the sixth month before the loan application
Income tax amount in the second quarter before the loan application	VAT amount in the second month before the loan application
Number of income tax payments in the past six months	VAT amount in the fourth month before the loan application
VAT amount in the third month before the loan application	Income tax amount in the first quarter before the loan application
VAT amount in the fifth month before the loan application	Number of income tax payments in the past six months
VAT amount in the sixth month before the loan application	Enterprise’s duration
Number of tax violations in the past three year	Total amount of income tax in the past six months

Number of tax violations in the past two year	VAT amount in the first month before the loan application
Number of tax violations in the past year	Number of tax violations in the past three year
	VAT amount in the fifth month before the loan application
	Registered capital
	Number of employees
	Income tax amount in the second quarter before the loan application
	Number of tax violations in the past two year
	Registration type = Private limited company
	Number of tax violations in the past six months

Source: Author.

6.1 SVM

This chapter still uses the Radial Basis Function kernel as the kernel function of SVM, and uses grid search and 5-fold cross-validation to optimize the parameters C and σ , the process is the same as that described in Chapter 5. The optimal parameters are shown in Table A42.

The analysis results in this part will be used to compare with the relevant results in Chapter 5, to clarify whether the index system optimized by the RF has better classification ability.

6.1.1 Training result of SVM classifier on tax features

The SVM classifier is trained on the new sample 3, the confusion matrix of the model is shown in Table A43, and the related performance indicators are shown in Table A44. As can be seen from the table, 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; 156 of the 161 non-defaulting enterprises can be correctly identified. Compared to the performance of the SVM classifier trained on sample 3 before feature optimization, which is shown in Table A25, the recall rate of the model remains the same at 89.47%, but the accuracy drops from 97.78% to 96.11%, and the performance seems to be declining. We further search for the point closest to the upper left corner on the ROC curve as shown in Figure B14, and find that the threshold corresponding to this point is 0.2473.

We reclassify the testing set by this threshold. Table A45 shows the confusion matrix after classification according to the optimal threshold, and Table A46 shows the relevant performance metrics. It can be found that after using the optimal threshold classification, the recall rate of the model increases from 89.47% to 94.74%, and the accuracy rate also increases from 96.11% to 97.22%. 157 out of 161 non-defaulting enterprises are accurately identified, and 18 out of 19 defaulting enterprises are accurately identified. The approval rate for the model rises from 86.74% to 87.78%, and the default rate falls from 1.27% to 0.63%. This is the highest

recall rate and the lowest default rate that can be achieved by using the SVM algorithm on all feature subsets so far, under the goal of balancing recall rate and accuracy rate, approval rate and default rate.

6.1.2 Training result of SVM classifier on tax features and basic features

The SVM classifier is trained on the new sample 5, and the confusion matrix of the model is shown in Table A47. Table A48 shows various performance metrics of the classifier on the testing set. It can be seen from the table that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; 160 of the 161 non-defaulting enterprises can be correctly identified. Compared with sample 5 before feature optimization, the recall rate of the model remains unchanged at 89.47%, but the accuracy rate increases from 97.78% to 98.33%, and the performance indicators are optimized. We further search for the point closest to the upper left corner on the ROC curve as shown in Figure B15, and find that the threshold corresponding to this point was 0.06259.

We reclassify the testing set according to this threshold. Table A49 shows the confusion matrix after classification according to the optimal threshold, and Table A50 shows the related performance metrics. It can be found that after using the optimal threshold classification, the recall rate of the model is still 89.47%, but the accuracy rate has been further improved, from 98.33% to 98.89%. All 161 non-defaulting enterprises are accurately identified, and 17 out of 19 defaulting enterprises are accurately identified. The approval rate for the model rises from 90% to 90.56%, while the default rate remains unchanged at 1.23%. It can be found that after using the optimal threshold classification, the model's identification accuracy of defaulting enterprises improved.

6.1.3 Performance comparison before and after feature index optimization

Figure 6.1 shows the performance difference between sample 3 and new sample 3, sample 5 and new sample 5 under the SVM algorithm. This comparison will reveal whether feature optimization helps to improve the classification performance. It is found that the use of RF for feature optimization is beneficial to improve the performance of SVM classifiers.

First, after feature optimization for sample 3, we extract 13 more important tax features to form new sample 3. The new data can improve the recall rate of the original model from 89.47% to 94.74%, and it can also keep the accuracy rate at the level of 97.22%. In the previous analysis, it is found that the use of sample 1, the basic data, for classification can achieve a recall rate of

100%, but the accompanying accuracy rate is only 65%, which is still not desirable for banks. The recall rate of 94.74% achieved by the new sample 3 is the highest recall rate that we can achieve by using the SVM algorithm under the goal of balancing the recall rate with the accuracy rate.

Secondly, after the feature optimization of sample 5, we extract 20 more important basic features and tax features to form a new sample 5. The recall rate of the SVM classifier trained on this data is consistent with the original model, both 89.47%, but the accuracy rate has been further improved to 98.89%, the approval rate has reached 90.56%, and the default rate is only 1.23%, which is much lower than the initial default rate of 10.6% in the testing set. This is also the highest accuracy rate and approval rate that the SVM learner can currently achieve for all subsets of data.

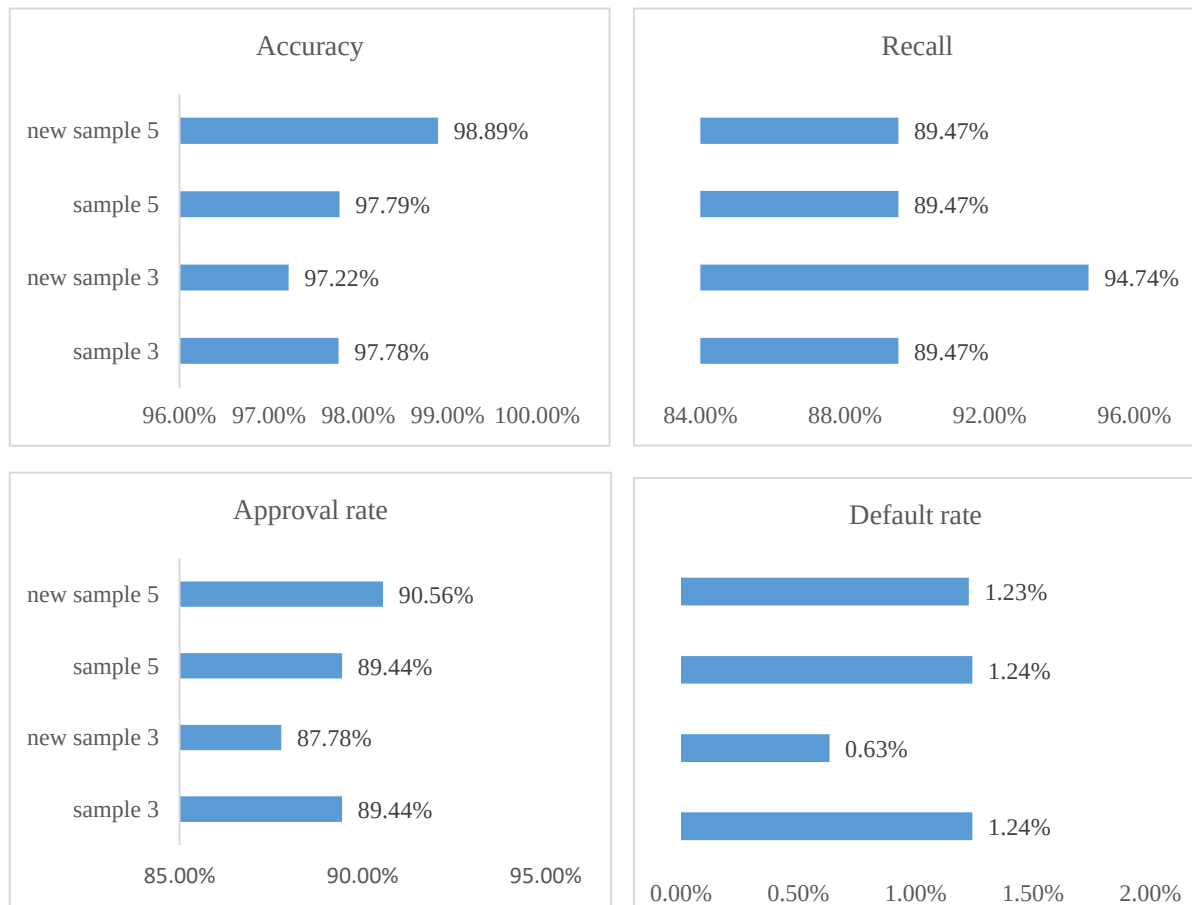


Figure 6.1 Performance comparison for samples before and after feature optimization

Source: Author.

6.2 LR

In this thesis, the L2-type regular term is added to the loss function of logistic regression. For the optimization of the hyperparameter C of the regular term, grid search and 5-fold cross-

validation method are used to solve the optimal parameter value under the goal of the highest average AUC. The results are shown in Table A51.

6.2.1 Training result of LR classifier on tax features

The LR classifier is trained on the new sample 3. The confusion matrix of the model and the related performance indicators are shown in Table A52 and Table A53. It can be seen that 17 of the 19 defaulting enterprises can be correctly identified by using the LR algorithm, with a recall rate of 89.47%; 160 of the 161 non-defaulting enterprises can be correctly identified, and the accuracy of the model reaches 98.33%, the approval rate is 90%, and the default rate is 1.23%.

Figure B16 shows the ROC curve of the new sample 3 under the LR algorithm. We further search for the point on the ROC curve that is closest to the upper left corner, and find that the threshold corresponding to this point is 0.64145. This thesis reclassifies the test set according to this threshold. Table A54 shows the confusion matrix after classification according to the optimal threshold, and Table A55 shows the related performance metrics. It can be found that after using the optimal threshold classification, the recall rate of the model remains unchanged at 89.47%, but the accuracy rate drops from 98.33% to 97.78%. Of the 161 non-defaulting firms, two are misclassified. This shows that the classifier with 0.64145 as the classification threshold does not perform as well as the classifier with 0.5 as the classification threshold. This may be due to the limited number of samples in the training set in this thesis, the predicted probability values of each sample are not completely continuous, and no enterprise has a probability value equal to 0.5, so there is a certain error when classifying according to the optimal threshold. Based on the above analysis, this thesis will abandon the optimal threshold classification results, and will still classify according to the original threshold, which is 0.5.

6.2.2 Training result of LR classifier on tax features and basic features

The LR classifier is trained on the new sample 5, and the confusion matrix of the model and related performance indicators are shown in Table A56 and Table A57. It can be seen that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; only 157 of the 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 96.67%. Under this classification algorithm, the approval rate of the sample reaches 88.33%, and the default rate is 1.26%.

This thesis also further solves the optimal classification threshold on the ROC curve, and reclassifies it according to the optimal threshold, but the classification results show that the

accuracy of the model further drops to 94.44%, and the recall rate remains unchanged at 89.47%. Therefore, classification according to the optimal threshold does not improve the performance index of the model. This thesis still chooses 0.5 as the classification threshold, and the classification results according to the optimal threshold are not shown here.

6.3 BPNN

The BPNN algorithm has many parameters, and this thesis mainly adjusts two important parameters: the number of hidden layers and neurons, and the regularization parameter α in the loss function. The grid search and 5-fold cross-validation method is still used to solve the optimal parameter values under the goal of the highest average AUC. The results are shown in Table A58.

The parameter tuning results show that for the new sample 3, the optimal number of hidden layers is 2, and the number of neurons is 100 and 50 respectively; while for the new sample 5, the optimal number of hidden layers is 2, and the number of neurons is 50 and 50 respectively.

6.3.1 Training result of BPNN classifier on tax features

The BPNN classifier is trained on the new sample 3, and the confusion matrix of the model and related performance indicators are shown in Table A59 and Table A60. It can be seen that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; 160 of the 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 98.33%. Under this classification algorithm, the approval rate of the sample reaches 90%, and the default rate is 1.23%.

This thesis solves the optimal classification threshold on the ROC curve, and reclassifies according to the optimal threshold, but the classification results show that the performance indicators of the model have not changed, and the accuracy and recall all have not changed. Therefore, the original classification threshold is still selected in this thesis, and the classification results by the optimal threshold are not shown here.

6.3.2 Training result of BPNN classifier on tax features and basic features

The BPNN classifier is trained on the new sample 5, and the confusion matrix of the model and related performance indicators are shown in Table A61 and Table A62. It can be seen from the tables that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of

89.47%; 159 of the 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 97.78%. Under this classification algorithm, the approval rate of the sample reaches 89.44%, and the default rate is 1.24%.

We also further solve the optimal classification threshold on the ROC curve, and reclassify it according to the optimal threshold. The classification results show that the recall rate of the model remains unchanged, but the accuracy rate has dropped from 97.78% to 96.67%. Therefore, this thesis will still use the original threshold classification results, and the optimal threshold classification results will not be shown here.

6.4 RF

The main parameter of the RF algorithm is the number of decision trees that called $n_estimators$. In this thesis, we use the grid search and 5-fold cross-validation method to solve the optimal number of decision trees under the goal of the highest average AUC. The results are shown in Table A63. The parameter tuning results show that for the new sample 3, the optimal number of decision trees is 19, and the new sample 5 is 26. In this thesis, the classifier will be trained and tested on the two datasets according to the optimal parameters.

6.4.1 Training result of RF classifier on tax features

The RF classifier is trained on the new sample 3, and the confusion matrix of the model and related performance metrics are shown in Table A64 and Table A65. It can be seen from Table A64 that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; 160 of the 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 98.33%. Under this classification algorithm, the approval rate of the sample reaches 90%, and the default rate is 1.23%.

We also further solve the optimal classification threshold on the ROC curve, and reclassify it according to the optimal threshold. The classification results show that the performance indicators of the model have not changed. Therefore, we will still use the original threshold classification results, and the optimal threshold classification results will not be shown here.

6.4.2 Training result of RF classifier on tax features and basic features

The RF classifier is trained on the new sample 5, and the confusion matrix of the model and related performance metrics are shown in Table A66 and Table A67. It can be seen from the

tables that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; all 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 98.8%. Under this classification algorithm, the approval rate of the sample reaches 90.56%, and the default rate is 1.23%.

Similarly, we further solve the optimal classification threshold on the ROC curve, and reclassify the testing set according to the optimal threshold. The classification results show that the performance indicators of the model have not changed. Therefore, we will still use the original threshold classification results, and the optimal threshold classification results will not be shown here

6.5 TSVM

The main parameters of the TSVM include the regular term coefficients C_u and C_l of the labeled samples and the unlabeled samples. The initial values of the two require $C_u \ll C_l$. In this thesis, the values of the two are set to be 0.001 and 1, respectively. RBF kernel is taken as the kernel function of the SVM, the parameter σ of the RBF kernel takes the default value, which is 1/the number of features.

6.5.1 Training result of TSVM classifier on tax features

The TSVM classifier is trained on the new sample 3, and the confusion matrix of the model and related performance indicators are shown in Table A68 and Table A69. It can be seen from Table A68 that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; 160 of the 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 98.33%. Under this classification algorithm, the approval rate of the sample reaches 90%, and the default rate is 1.23%.

In this thesis we further solve the optimal classification threshold on the ROC curve, and reclassify the testing set according to the optimal threshold. The classification results show that the performance indicators of the model have not changed. Therefore, we will still use the original threshold classification results, and the optimal threshold classification results will not be shown here. In addition, it should be noted that this thesis uses the TSVM classifier to predict the non-passing enterprises, and find that 1,304 non-passing enterprises in the sample are classified as “will not default”.

6.5.2 Training result of TSVM classifier on tax features and basic features

The TSVM classifier is trained on the new sample 5, and the confusion matrix of the model and related performance indicators are shown in Table A70 and Table A71. It can be seen from the tables that 17 of the 19 defaulting enterprises can be correctly identified, with a recall rate of 89.47%; all 161 non-defaulting enterprises have been correctly identified. The accuracy of the model is 98.89%. Under this classification algorithm, the approval rate of the sample reaches 90.56%, and the default rate is 1.23%.

Similarly, we further solve the optimal classification threshold, and reclassify the testing set according to the optimal threshold. The classification results show that the performance indicators of the model have not changed. Therefore, we will still use the original threshold classification results, and the optimal threshold classification results will not be shown here. In this thesis, the TSVM classifier is used to predict the non-passing enterprises, and it is found that 1304 non-passing enterprises in the sample are classified as “will not default”.

6.6 Performance comparison of classifiers trained by different algorithms

Figures 6.2 show the performance of the classifiers trained by SVM, LR, BPNN, RF and TSVM on the new sample 3, which contains the tax features. These five algorithms include supervised and semi-supervised algorithms, individual algorithms and ensemble algorithms. If the performance of an algorithm is improved after using the optimal threshold classification on the ROC curve, we use the optimal threshold for classification; if the performance indicators remain unchanged or deteriorate after using the optimal threshold classification, the original threshold standard of the model will be used for classification. For example, the original threshold of the LR algorithm is 0.5, and the original threshold of SVM and TSVM is 0.

By comparing the various indicators in Figure 6.2, the following findings are made.

First, four of the five algorithms have consistent performance, namely LR, BPNN, RF, and TSVM. These four algorithms can achieve a classification accuracy of 98.33%, a recall rate of 89.47% and a false positive rate of 0.62%, which means that these classifiers will misjudge 0.62% of non-defaulting enterprises as defaulting ones, however, 89.47% of enterprises that will default can be identified. If the classifier is used to re-approve the testing set samples, the approval rate of 90% can be achieved, and the default rate will drop from 10.56% to 1.23%. Since these four algorithms have consistent classification performance under the current threshold, we can choose the best one by comparing their AUC values. The AUC value

comprehensively reflects the classification ability of the algorithm at different thresholds. As shown in the figure, the AUC value of the TSVM algorithm is 97.14%, which is the highest among the four. It can be considered that the classification performance of the TSVM algorithm at the current threshold is consistent with LR, BPNN, and RF, but the comprehensive classification ability at all thresholds is stronger than that of LR, BPNN, and RF.

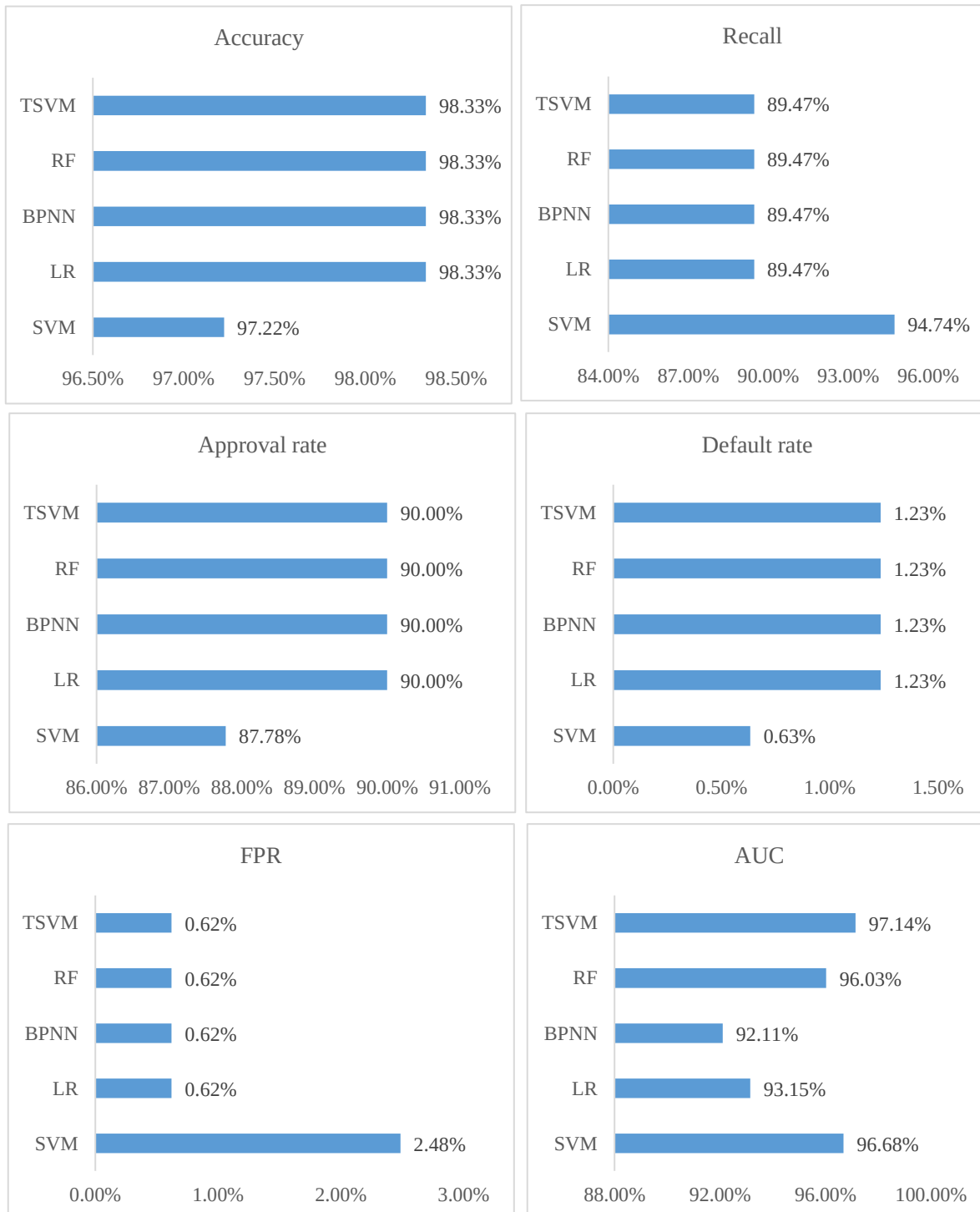


Figure 6.2 Performance comparison of five algorithms on the new sample 3

Source: Author.

Second, there are differences in the performance of the SVM algorithm compared to the TSVM algorithm. The SVM algorithm has a higher recall rate and can identify 94.74% of defaulting enterprises, but the accuracy rate is 97.22%, which is lower than 98.33% of TSVM; the false positive rate also reaches 2.48%, which is higher than 0.62% of TSVM. It can be seen that the SVM algorithm uses stricter standards and lower thresholds to classify more enterprises as “will default” to improve the recall rate, and accidentally rejects more non-defaulting enterprises. Comparing the AUC values of TSVM and SVM, it can be found that the AUC value of TSVM is higher than that of SVM. Therefore, in these two classification algorithms, TSVM will still be selected in this thesis.

Figure 6.3 shows the performance of the classifiers trained by the five algorithms on new samples 5, which contains the basic features and tax features. The following findings are found from the figure.

First, among the five algorithms, TSVM, RF and SVM have consistent performance under the current threshold and are the best. All three algorithms can achieve a classification accuracy of 98.89%, and achieve a recall rate of 89.47% under the premise that the FPR is 0. This means that these three algorithms can identify 89.47% of the “will default” enterprises without misjudging a “non-defaulting enterprise”. If the classifier is used to re-approve the testing set samples, the approval rate of 90.56% can be achieved, and the default rate will drop from 10.56% to 1.23%. At the same time, we can find that the classification performance of BPNN and LR is worse than that of TSVM, RF and SVM. Although the recall rates of the five are 89.47%, the classification accuracy of BPNN is 97.88%, while LR is only 96.67%. This results in lower approval rates, higher default rates, and higher FPRs for both. Therefore, at the current threshold, we can rank the performance of the five algorithms: $LR < BPNN < SVM = TSVM = RF$.

Second, since the classification performance of TSVM, RF and SVM under the current threshold is consistent, we will select the optimal algorithm by comparing the comprehensive performance of the classification performance under different thresholds, that is, the AUC value. As can be seen from the figure, the AUC value of RF reaches 96.7%, which is the highest among the three. Therefore, we believe that the RF algorithm has better comprehensive classification performance on new sample 5.

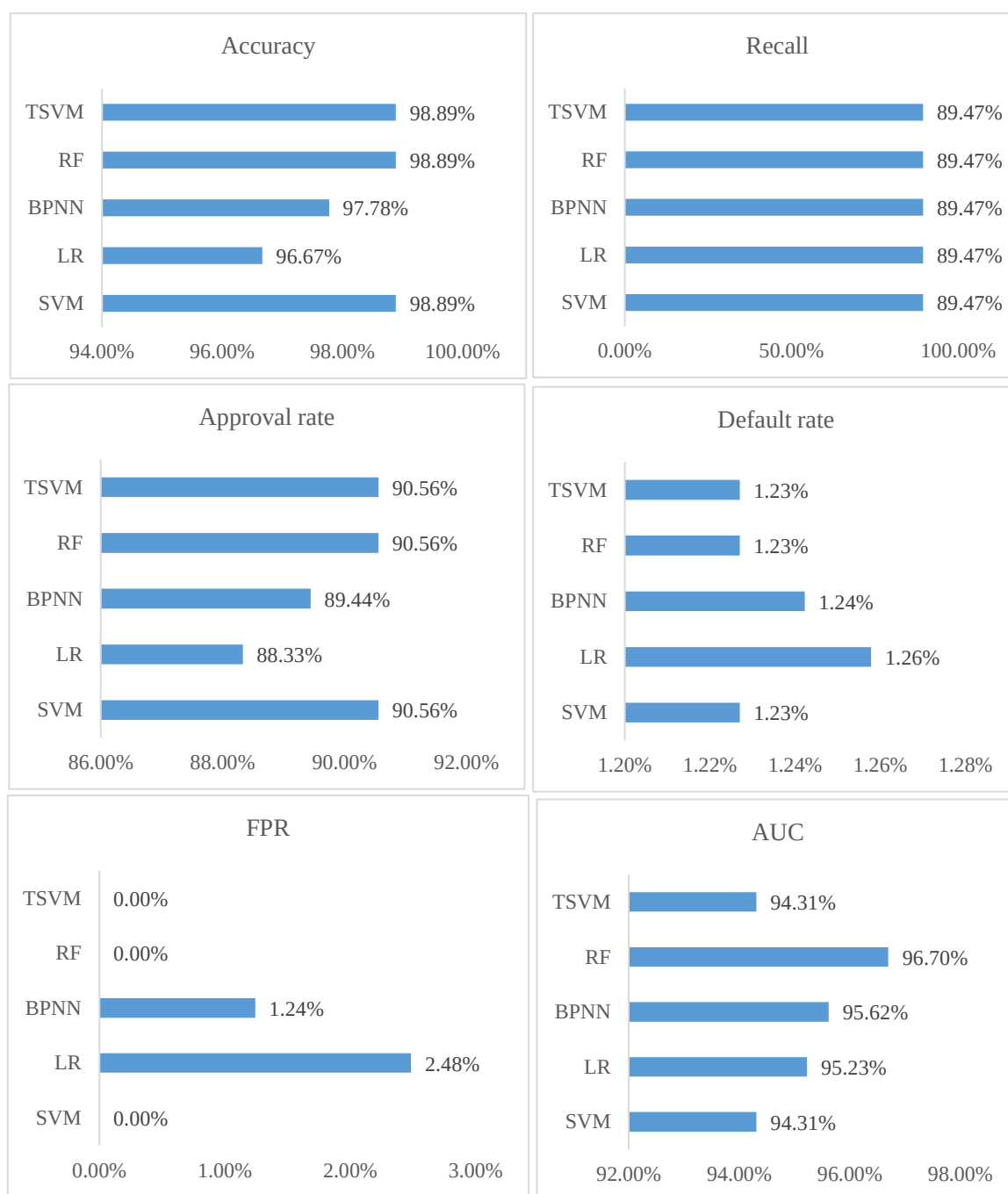


Figure 6.3 Performance comparison of five algorithms on the new sample 5

Source: Author.

6.7 Chapter summary

Based on the optimized new sample 3 (tax features) and new sample 5 (basic features and tax features), this chapter uses five algorithms of supervised learning and semi-supervised learning to train and test the classifier. Among them, the supervised learning algorithm includes the individual algorithm such as SVM, LR and BPNN, and the ensemble algorithm such as RF. At

the same time, we use the semi-supervised learning algorithm TSVM. Through the analysis, the following findings are found.

First, using RF to optimize features and remove unimportant features can effectively improve the classification ability of feature data. In this thesis, the SVM algorithm is used to train the feature data before and after optimization. It is found that after optimizing sample 3 (that is, the tax features), reducing the 22 features to 13 features can improve the recall rate of the SVM classifier from 89.47% to 94.74%, while the accuracy rate remains unchanged at 97.22%. After optimizing sample 5 (basic features and tax features), reducing the 32 features to 20 features can improve the accuracy rate of the SVM classifier from 97.78% to 98.89%, while the recall remains the same.

Second, the classification performance of the ensemble algorithm and the semi-supervised algorithm is better than that of the individual algorithm. On the new sample 3, the semi-supervised algorithm TSVM has the best performance, while on the new sample 5, the ensemble algorithm RF has the best classification performance, which is better than the individual algorithms such as SVM, BPNN and LR. Although on the new sample 3, the performance of TSVM is consistent with that of RF, LR and BPNN under the current classification threshold, but the comprehensive classification ability under different thresholds is still stronger.

Why does the TSVM algorithm not significantly improve the accuracy rate or recall rate of the model after adding the feature data of the non-passing enterprises in the classifier training process? There maybe two reasons. First, the non-passing samples do not provide more feature distribution information. After verification, the predicted labels of the 1304 non-passing samples collected are all “will not default” when classified by TSVM. On the one hand, it shows that the original loan review rule of “Tax E Financing” is unreasonable, and it is highly misjudged for enterprises that fail to pass. On the other hand, it also shows that non-passing samples do not provide information about the characteristics of “defaulting” enterprises. Adding the non-passing samples only increases the information of the majority category in the sample, and cannot provide more information for identifying the minority category, that is, the defaulting enterprises. The second is that there is a certain amount of information noise in the sample data of this thesis, that is, in the current information set, there are default enterprises that cannot be identified by the optimization algorithm, which is the upper limit of performance determined by the data.

Third, compared with the optimized tax feature set (new sample 3), the optimized basic feature and tax feature set (new sample 5) have better classification performance. On the optimized tax feature set, the TSVM classifier has the best performance among the five

algorithms, as it can achieve a recall rate of 89.47%, an accuracy rate of 98.33%, and an FPR of 0.62%. While on the optimized basic feature and tax feature set, the RF classifier has the best performance, achieving a recall rate of 89.47%, an accuracy rate of 98.89% and an FPR of 0%. The comparison shows that the recall rates of the two optimal classifiers are the same, but the RF classifiers on the basic feature and tax feature set have higher classification accuracy and lower FPR. Therefore, for the credit risk evaluation of “Tax E Financing”, this thesis believes that using the optimized 20 basic features and tax features as the index system and using the RF algorithm for classifier training can have the best performance.

Chapter 7: Conclusions, Suggestions and Limitations

7.1 Research conclusions

In this thesis, we attempt to improve the loan review mechanism of “Tax E Financing”, including the construction and optimization of credit risk evaluation index system and the selection of credit risk evaluation method. To this end, this thesis takes enterprises applying for loans from “Tax E Financing” as a sample, collects and calculates the basic features, financial features and tax features of defaulting enterprises, non-defaulting enterprises and non-passing enterprises. Then, based on the above feature data and the default behavior of enterprises, the following analysis is carried out, and relevant conclusions are drawn.

First, we compare and analyze the features of defaulting enterprises, non-defaulting enterprises and non-passing enterprises, and have the following two findings. Firstly, in terms of basic features, defaulting enterprises have low loan application amount and short business duration; in terms of financial features defaulting enterprises have higher accounts receivable, prepaid accounts, other receivables and inventories, and higher construction in progress, intangible assets, long-term deferred expenses, various liabilities, financial expenses, non-operating income and profits, but their financial data is unreliable; in terms of tax features, both the amount and frequency of income tax and VAT payments of defaulting enterprises are significantly lower. Secondly, some features of the non-passing enterprises are better than those of the non-defaulting enterprises, such as the loan application amount, the number of employees, and the amount of VAT and income tax paid in the past six months, indicating that they have a high probability of having been rejected by mistake.

Second, we divide the feature dataset into 7 subsamples, namely “basic features”, “financial features”, “tax features”, “basic features + financial features”, “basic features + tax features”, and “basic features + financial features + tax features”. Then SVM is used to are used to train classifier on each subset. Comparing the performance of each classifier, it is found that “tax features” and “basic features + tax features” have the best classification performance. “Financial feature” has the worst performance index. Mixing “financial features” with “basic features”, “tax features” and “basic features + tax features” will not only fail to play a positive role in the distinction between defaulting enterprises and non-defaulting enterprises, but it will

have a negative impact. At the same time, we use the RF algorithm to optimize the “basic features” and “basic features + tax features” set. It is found that removing features whose importance is lower than the threshold can effectively improve the classification ability of feature data.

Third, we use various algorithms of supervised learning and semi-supervised learning to train and test the classifier based on the optimized “tax features” and “basic features + tax features”. Among them, supervised learning algorithms include individual algorithms and ensemble algorithms. Individual algorithms used in this thesis include SVM, LR and BPNN. Meanwhile, ensemble algorithm RF and semi-supervised learning algorithm TSVM are used. The study finds that for the credit risk evaluation of “Tax E Financing”, using the optimized 20 basic features and tax features as the index system, using the RF algorithm for classifier training can have the best performance, and the accuracy rate of the classifier reaches 98.89 %, while the FPR is 0, the recall rate reaches 89.47%.

7.2 Policy suggestions

According to the conclusions of Chapter 4, Chapter 5 and Chapter 6, we know that the tax features have a very high ability to identify the credit risk of enterprises. Therefore, the following suggestions are made for banks’ credit risk assessment of small and micro enterprises and the government’s “Interaction between bank and tax” work.

First, banks should abandon the “one-vote veto” system and adopt a more reasonable credit risk evaluation mechanism when evaluating the credit risk of small and micro enterprises. Firstly, when building the credit risk evaluation index system of small and micro enterprises, their financial characteristics should be abandoned, and more reliable and real basic characteristics and tax characteristics should be fully used. Secondly, banks should make full use of machine learning algorithms to deeply mine data features, build classification models with higher performance, improve the accuracy of credit risk prediction models, and achieve a balance between high approval rate and low default rate. Finally, based on the absolute reliability and classification validity of tax features banks can also introduce tax features into the index system when assessing the credit risk of large and medium-sized enterprises.

Second, the government should formulate policies to allow tax authorities to more fully share relevant tax information with banks, and at the same time introduce policies to encourage banks to use tax information more fully and optimize their loan approval mechanisms. Firstly, based on the effectiveness of tax information for enterprise credit risk assessment, banks should

further expand the scope of tax information sharing, from small and micro enterprises to medium and large enterprises, and comprehensively improve the application scope of tax information in enterprise credit risk assessment activities. Secondly, considering the importance of “Tax E Financing” and other products in solving the financing problems of small and micro enterprises, the government can provide certain tax relief for banks according to the number of enterprises and loan amount covered by the relevant products and loan services of each bank. It can force banks to make more in-depth use of tax information, optimize the loan approval mechanism, and improve the loan approval rate. At the same time, it can also effectively encourage small and micro enterprises to standardize their taxation behavior and contribute to the improvement of national taxation.

7.3 Research limitations

The limitations of this thesis are mainly reflected in the sample data.

First, due to data limitations, we failed to collect the relevant information of all loan application enterprises since the establishment of “Tax E Financing”. Only the data of the enterprises that applied for loans in the second half of 2017 and the defaulting enterprises that applied for loans in 2018 and 2019 were collected.

Second, due to data limitations, we cannot collect the business owner information of each sample enterprise, including their age, gender, credit status.

Third, due to the strict loan review mechanism of “Tax E Financing”, the number of defaulting enterprises is small, and the imbalance of sample data is serious. To a certain extent, it is not conducive to the model to effectively identify the characteristics of defaulting enterprises. In future research, if more sample data can be collected, business owner information can be collected, and the number of defaulting enterprises can be further increased, it will help to further optimize the “Tax E Financing” credit risk evaluation index system, and train a classification model with higher performance.

[This page is deliberately left blank.]

Bibliography

- Aharony, J., Jones, C. P., & Swary, I. (1980). An analysis of risk and return characteristics of corporate bankruptcy using capital market data. *Journal of Finance*, 35(4), 1001-1016.
- Ahlin, C., & Townsend, R. M. (2007). Using repayment data to test across models of joint liability lending. *The Economic Journal*, 117(517), F11-F51.
- Akerlof, G. A. (1970). The market for “lemons” : Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488-500.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589-609.
- Altman, E. I. (1973). Predicting railroad bankruptcies in America. *The Bell Journal of Economics Management Science*, 4(1), 184-211.
- Altman, E. I., & Brenner, M. (1981). Information effects and stock market response to signs of firm deterioration. *Journal of Financial & Quantitative Analysis*, 16(1), 35-51.
- Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance*, 1(1), 29-54.
- Altman, E. I., & Loris, B. (1976). A financial early warning system for over-the - counter broker - dealers. *The Journal of Finance*, 31(4), 1201-1217.
- Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, 18(3), 505-529.
- Altman, E. I., & Sabato, G. (2007). Modeling credit risk for SMEs: Evidence from the US market. *A Journal of Accounting, Finance and Business Studies*, 3, 332-357.
- Alzamora, G. S., Aceituno-Rojas, M. R., & Condori-Alejo, H. I. (2022). An assertive machine learning model for rural micro credit assessment in Peru. *Procedia Computer Science*, 202, 301-306.
- Arene, C. (1992). Loan repayment and technical assistance among small-holder maize farmers in Nigeria. *African Review of Money Finance & Banking*, 1, 63-74.
- Aziz, A., Emanuel, D. C., & Lawson, G. H. (1988). Bankruptcy prediction - an investigation of cash flow based models [1]. *Journal of Management Studies*, 25(5), 419-437.
- Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly*, 26(3), 243-268.
- Bai, S. B. (2010). 基于有序 logistic 模型的企业供应链融资风险预警研究 [A research into the risk early - warning of enterprise supply chain financing based on ordered logistic model]. *Economic Systems*, (6), 66-71.
- Bai, S. Z., & Li, S. (2013). 基于 BP 神经网络的供应链金融风险评估研究 [Supply chain finance risk evaluation research based on BP neural network]. *Commercial Research*, (1), 27-31.
- Barasinska, N., & Schäfer, D. (2014). Is crowdfunding different? Evidence on the relation between gender and funding success from a German peer-to-peer lending platform. *German Economic Review*, 15(4), 436-452.
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71-111.
- Beaver, W. H. (1968). Market prices, financial ratios, and the prediction of failure. *Journal of Accounting Research*, 6(2), 179-192.

- Behr, P., & Guettler, A. (2007). Credit risk assessment and relationship lending: An empirical analysis of German small and medium - sized enterprises. *Journal of Small Business Management*, 45(2), 194-213.
- Bei, Y. H., Yang, L., & Wang, Y. H. (2012). 汽车制造业在供应链金融模式下的信用风险评估 [Credit risk evaluation of car - making industry under supply chain financing model]. *Logistics Technology*, (31), 379-382.
- Bennardo, A., Pagano, M., & Piccolo, S. (2015). Multiple bank lending, creditor rights, and information sharing. *Review of Finance*, 19(2), 519-570.
- Berger, A. N., & Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *The Economic Journal*, 112(477), F32-F53.
- Berger, S. C., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *Business Research*, 2(1), 39-65.
- Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, 12(1), 1-25.
- Bollazzi, F., Risalvato, G., & Venezia, C. (2019). Asymmetric information and deal selection: Evidence from the Italian venture capital market. *International Entrepreneurship and Management Journal*, 15(3), 721-732.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (2017). *Classification regression trees*. Routledge.
- Brown, K., & Moles, P. (2014). *Credit risk management*. Mc Graw Hill Professional.
- Brown, M., Jappelli, T., & Pagano, M. (2009). Information sharing and credit: Firm-level evidence from transition countries. *Journal of Financial Intermediation*, 18(2), 151-172.
- Cao, C., & Zhang, M. (2022). Credit Risk Evaluation of Quantum Communications Listed Companies in China Based on Fermatean Fuzzy TOPSIS. *Procedia Computer Science*, 199, 361-368.
- Caouette, J. B., Altman, E. I., Narayanan, P., & Nimmo, R. (1998). *Managing credit risk: the next great financial challenge*. Wiley.
- Casey, C., & Bartczak, N. (1985). Using operating cash flow data to predict financial distress: Some extensions. *Journal of Accounting Research*, 23(1), 384-401.
- Castillo, J. A., Mora-Valencia, A., & Perote, J. (2018). Moral hazard and default risk of SMEs with collateralized loans. *Finance Research Letters*, 26(9), 95-99.
- Castillo, J. A., Mora-Valencia, A., & Perote, J. (2021). Moral hazard index for credit risk to SMEs. *International Economics*, In press.
- Chang, A.-Y., Whitehouse, D. J., Chang, S.-L., & Hsieh, Y.-C. (2001). An approach to the measurement of single-machine flexibility. *International Journal of Production Research*, 39(8), 1589-1601.
- Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73, 914-920.
- Charitou, A., Neophytou, E., & Charalambous, C. (2004). Predicting corporate failure: empirical evidence for the UK. *European Accounting Review*, 13(3), 465-497.
- Chen, D., Li, X., & Lai, F. (2017). Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China. *Electronic Commerce Research*, 17(4), 553-583.
- Chen, L., Xie, Y., Li, P., & Li, Q. (2019). 借款陈述文字中的违约信号——基于 P2P 网络借贷的实证研究 [The signal of default risk from the description-text based on the empirical research of P2P lending]. *Chinese Journal of Management Science*, 27(4), 37-47.

- Chen, S., Huang, R., & Yuan, Y. (2010). KMV 模型在我国上市公司信用风险评级中的应用 [Application of KMV model in credit risk rating of listed companies in China]. *Modern Business*, (2), 27.
- Chen, X., Huang, B., & Ye, D. (2018). The role of punctuation in P2P lending: Evidence from China. *Economic Modelling*, 68, 634-643.
- Chen, Y. F. (2020). 企业战略与商业信用 [Enterprise strategy and business credit]. China Renmin University Press.
- Cheng, C., & Yang, L. (2022). What drives the credit constraints faced by Chinese small and micro enterprises? *Economic Modelling*, 113, 105898.
- Chung, H. M., & Gray, P. (1999). Special section: Data mining. *Journal of Management Information Systems*, 16(1), 11-16.
- Claessens, S., Feijen, E., & Laeven, L. (2008). Political connections and preferential access to finance: The role of campaign contributions. *Journal of Financial Economics*, 88(3), 554-580.
- Clark, T. A., & Weinstein, M. I. (1983). The behavior of the common stock of bankrupt firms. *The Journal of Finance*, 38(2), 489-504.
- Craven, M. W., & Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, 13(2-3), 211-229.
- Crosbie, P., & Bohn, J. (2019). Modelling default risk. In C. Michel, G. Dan, & W. Zvi (Eds.), *World scientific reference on contingent claims analysis in corporate finance: Volume 2: Corporate debt valuation with CCA* (pp. 471-506). World Scientific.
- Crouhy, M., Galai, D., & Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking and Finance*, 24(1-2), 59-117.
- Dai, H., Ding, G., & Wu, X. (2016). KMV 模型在上市公司信用风险管理中的应用研究 [Application research of KMV Model in credit risk management of listed companies]. *Journal of Development - Oriented Finance Research*, 8(4), 37-45.
- Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(1), 167-179.
- Deakin, E. B. (1977). Business failure prediction: An empirical analysis. In E. I. Altman & A. W. Sametz (Eds.), *Financial crises: institutions and markets in a fragile environment* (pp. 72-98). Wiley.
- Deng, A., Xiong, J., & Zhang, F. (2010). Order financing risk pre-warning model based on BP network. *Journal of Intelligence*, 29, 23-28.
- Ding, J., Huang, J., Li, Y., & Meng, M. (2019). Is there an effective reputation mechanism in peer-to-peer lending? Evidence from China. *Finance Research Letters*, 30, 208-215.
- Djankov, S., McLiesh, C., & Shleifer, A. (2007). Private credit in 129 countries. *Journal of Financial Economics*, 84(2), 299-329.
- Dong, R. (2011). 基于 Logit 模型的微型企业信用评价研究 [Research on credit evaluation of micro enterprises based on Logit model]. *New Finance*, 8(270), 54-57.
- Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending - Evidence from two leading European platforms. *Journal of Banking & Finance*, 64, 169-187.
- Doumpos, M., Niklis, D., Zopounidis, C., & Andriosopoulos, K. (2015). Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms. *Journal of Banking & Finance*, 50, 599-607.
- Duan, W., & Lu, X. (2015). 基于代价敏感支持向量机的银行信用风险评估模型 [Evaluation model of bank's credit risk based on cost sensitive support vector machine]. *Journal of Jiangxi Science & Technology Normal University*, (6), 75-79.
- Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-

- peer lending. *The Review of Financial Studies*, 25(8), 2455-2484.
- Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial & Quantitative* 7(2), 1477-1493.
- Edmister, R. O., & Schlarbaum, G. G. (1974). Credit policy in lending institutions. *Journal of Financial & Quantitative Analysis*, 9(3), 335-356.
- Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1), 54-70.
- Fan, A., & Palaniswami, M. (2000, July 24-27). *Selecting bankruptcy predictors using a support vector machine approach*. International Joint Conference on Neural Networks, Como, Italy.
- Fan, N. (2002). Creditmetric 模型及其对我国银行信用卡信用风险管理的借鉴 [Credit metric model and its application in credit risk management by China's banks]. *Financial Forum*, 7(5), 50-54.
- Fang, K., Wu, J., Zhu, J., & Xie, B. (2010). 信贷信息不对称下的信用卡信用风险研究 [Forecasting of credit card credit risk under asymmetric information based on nonparametric random forests]. *Economic Research Journal*, 7(5), 50-54.
- Fang, K., Zhang, G., & Zhang, H. (2014). 基于 Lasso-logistic 模型的个人信用风险评价方法 [Individual credit risk prediction method: Application of Lasso - logistic model]. *The Journal of Quantitative & Technical Economics*, 2(1), 125-136.
- Fantazzini, D., & Figini, S. (2009a). Default forecasting for small-medium enterprises: Does heterogeneity matter? *International Journal of Risk Assessment and Management*, 11(1-2), 138-163.
- Fantazzini, D., & Figini, S. (2009b). Random survival forests models for SME credit risk measurement. *Methodology and Computing in Applied Probability*, 11(1), 29-45.
- Fidrmuc, J., & Hainz, C. (2009). Default rates in the loan market for SMEs: Evidence from Slovakia. *Economic Systems*, 34(2), 133-147.
- Fitzpatrick, P. J. (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies. *Análise Molecular Do Gene Wwox*, 3(2), 598-605.
- Fonseca, D. P., Wanke, P. F., & Correa, H. L. (2020). A two-stage fuzzy neural approach for credit risk assessment in a Brazilian credit card company. *Applied Soft Computing*, 92, 106329.
- Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: Lessons from peer-to-peer lending. *International Journal of Industrial Organization*, 51, 185-222.
- Fu, Q., & Li, Y. (2005). Logistic 模型在上市公司信用风险评价中的应用 [Application of Logistic model in credit risk assessment of listed companies]. *Journal of Chongqing Architecture University*, (5), 116-119, 125.
- Galindo, A., & Miller, M. J. (2001, September 19). *Can credit registries reduce credit constraints? Empirical evidence on the role of credit registries in firm investment decisions*. Annual Meetings of the Inter - American Development Bank, Santiago, Chile.
- Garen, J. E. (1994). Executive compensation and principal-agent theory. *Journal of Political Economy*, 102(6), 1175-1199.
- Gentry, J. A. (1985). Classifying Bankrupt firms with funds flow component. *Journal of Accounting Research*, (23), 146-160.
- Gentry, J. A., Newbold, P., & Whitford, D. T. (1984). Bankruptcy, working capital and funds flow components. *Managerial Finance*, 10(3), 26-29.
- Gordy, M. B. (2002). Saddle point approximation of credit risk. *Journal of Bank and Finance*, 26(7), 1335-1353.
- Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3), 22-45.
- Greiner, M. E., & Wang, H. (2010). Building consumer-to-consumer trust in e-finance

- marketplaces: An empirical analysis. *International Journal of Electronic Commerce*, 15(2), 105-136.
- Guo, W., Chen, Z., & Zhong, M. (2013). 基于 MLP 神经网络构建小企业信用风险预警模型 [Construction of early warning model of small business credit risk based on MLP neural network]. *Finance and Accounting Monthly*, (6), 22-26.
- Gupton, G., Finger, C., & Bhatia, M. (1997). *CreditMetrics Technical Document*. JP Morgan & Co.
- Han, L., & Han, L. (2017). 正交支持向量机及其在信用评分中的应用 [Orthogonal support vector machine and its application in credit scoring]. *Journal of Industrial Engineering/Engineering Management*, 31(2), 128-136.
- Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, 48(SPL), 138-S149.
- Hou, W., Wang, X., Zhang, H., Wang, J., & Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems*, 208, 106462.
- Hsieh, W. W. (2009). *Machine learning methods in the environmental sciences: Neural networks and kernels*. Cambridge university press.
- Hu, H., Zhang, L., & Zhang, D. (2012). 供应链金融视角下的中小企业信用风险评估研究——基于 SVM 与 BP 神经网络的比较研究 [Research on SMEs credit risk assessment from the perspective of supply chain finance – a comparative study on the SVM model and bp model]. *Management Review*, 24(11), 70-80.
- Huo, L., Song, R., & Shen, Z. (2011). KMV 模型在农业上市公司信用风险度量中的应用 [Application of KMV model in credit risk measurement of agricultural listed companies]. *Finance and Accounting Monthly*, (5), 56-58.
- Ifeanyi, A. O., & Blessing, C. O. (2012). Economic analysis of loan repayment capacity of smallholder cooperative farmers in Yewa North Local Government Area of Ogun State, Nigeria. *African Journal of Agricultural Research*, 7(13), 2051-2062.
- Jaffee, D. M., & Russell, T. (1976). Imperfect information, uncertainty, and credit rationing. *The Quarterly Journal of Economics*, 90(4), 651-666.
- Jappelli, T., & Pagano, M. (2002). Information sharing, lending and defaults: Cross – country evidence. *Journal of Banking & Finance*, 26(10), 2017-2045.
- Jia, S. H., & Shi, Y. J. (2003). 商业银行的中小企业信贷风险因素及其管理对策 [The credit risk factors and management countermeasures of SMEs in commercial banks]. *Journal of Zhejiang University (Humanities and Social Sciences)*, (2), 57-64.
- Jiang, C. Q., Wang, R. Y., & Ding, Y. (2017). 融入软信息的 P2P 网络借贷违约预测方法 [A P2P network loan default prediction method incorporating soft information]. *Management Science in China*, 25(11), 12-21.
- Jin, X., Ke, Y., & Chen, X. (2022). Credit pricing for financing of small and micro enterprises under government credit enhancement: Leverage effect or credit constraint effect. *Journal of Business Research*, 138, 185-192.
- Joachims, T. (1999, June 27-30). *Transductive inference for text classification using support vector machines*. International Conference on Machine Learning Beld, Slovenia. Yugoslavia.
- Jones, D., & Mingo, J. (1999). Credit risk modeling and internal capital allocations: implications for a models – based regulatory Capital Standard. *Journal of Economics and Business*, 51(2), 79-108.
- Karlan, D. S. (2007). Social connections and group banking. *The Economic Journal*, 117(517), F52-F84.

- Kartasheva, A. V., & Traskin, M. (2011). *Insurers' insolvency prediction using random forest classification*. [Doctoral dissertation], The Wharton School, University of Pennsylvania.
- Kautonen, T., Fredriksson, A., Minniti, M., & Moro, A. (2020). Trust-based banking and SMEs' access to credit. *Journal of Business Venturing Insights*, 14, e00191.
- Keasey, K. A., & Watson, R. (1987). Non - Financial Symptoms and the Prediction of Small Company Failure: A Test of Argenti's Hypotheses. *Journal of Business Finance and Accounting*, 14(3), 335-354.
- Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural computation*, 15(7), 1667-1689.
- Klein, D. B. (1992). Promise keeping in the great society: a model of credit information sharing. *Economics and Politics*, 4(2), 117-136.
- Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing*, 107, 107391.
- Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1), 19-37.
- 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.
- Li, A. X. (2014). 社会资本对农户信用行为影响的机理分析 [Mechanism analysis of the influence of social capital on farmers' credit behavior]. *Collected Essays on Finance and Economics*, (1), 49-55.
- Li, B., Li, X., & Zhao, Z. (2006). Novel algorithm for constructing support vector machine regression ensemble. *Journal of Systems Engineering and Electronics*, 17(3), 541-545.
- Li, H. (2012). *统计学习方法 [Statistical learning method]*. Tsinghua University Press.
- Li, J., & Hu, J. (2019). Does university reputation matter? Evidence from peer-to-peer lending. *Finance Research Letters*, 31, 66-77.
- Li, Q. (2018a). 社会网络对贫困地区农户信贷违约风险的影响:抑制还是激励? [The impact of social network on credit default risk of rural households in poverty-stricken areas: inhibition or incentive?] *China Rural Survey*, (5), 45-66.
- Li, Q. (2018b). 社会资本对农户信贷违约影响的机制分析 [Mechanism analysis of the impact of social capital on farmers' credit default]. *Agricultural Technology Economy*, (2), 104-118.
- Li, S., Zhou, X., & Bao, Z. (2018). 基于支持向量机的商业银行对中小信贷企业选择方法的研究 [Research of loan enterprise selection for bank based on support vector machine]. *Mathematics in Practice and Theory*, 48(11), 299-305.
- Liberti, J. M., & Petersen, M. A. (2019). Information: Hard and soft. *Review of Corporate Finance Studies*, 8(1), 1-41.
- Lin, C., & Peng, G. (2007). 随机森林在企业信用评估指标体系确定中的应用 [Application of random forest on selecting evaluation index system for enterprise credit assessment]. *Journal of Xiamen University (Natural Science)*, 46(2), 199-203.
- Lin, H.-T., & Lin, C.-J. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *Neural Computation*, 3(1-32), 16.
- Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1), 17-35.
- Lin, S. L. (2009). A new two - stage hybrid approach of credit risk in banking industry. *Expert Systems with Applications*, 36(4), 8333-8341.

- Liu, C., Xie, J., Zhao, Q., Xie, Q., & Liu, C. (2019). Novel evolutionary multi-objective soft subspace clustering algorithm for credit risk assessment. *Expert Systems with Applications*, 138, 112827.
- Lo, S. L., Chiong, R., & Cornforth, D. (2015). Using support vector machine ensembles for target audience classification on twitter. *PLoS ONE*, 10(4), 1-20.
- Longstaff, F. A., & Schwartz, E. S. (1995). A Simple Approach to Valuing Risky Fixed and Floating Rate Debt. *The Journal of Finance*, 50(3), 789-819.
- Lu, H., Wang, B., Wang, H., & Zhao, T. (2020). Does social capital matter for peer-to-peer-lending? Empirical evidence. *Pacific-Basin Finance Journal*, 61, 101338.
- Lu, S., & Ooyang, Z. (1999). 商业银行信用风险跟踪预警监测模型 [Commercial bank credit risk tracking and early warning model]. *The Journal of Quantitative & Technical Economics*, 1(1), 59-62.
- Lussier, R. N., & Pfeifer, S. (2001). A crossnational prediction model for business success. *Journal of Small Business Management*, 39(3), 228-239.
- Ma, R. (2006). KMV 模型运用于中国上市公司财务困境预警的实证检验 [Testing KMV on the Financial Distress of Listed Companies in China]. *Application of Statistics and Management*, (5), 593-601.
- Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, 200, 116889.
- Maddala, G. S. (1983). *Limited - dependent and qualitative variables in econometrics*. Cambridge University Press.
- Majić, T., Pongrac, B., & Richter, G. (2015). Information asymmetry and moral hazard in financial economics. *Tehnički Glasnik*, 9(2), 209-215.
- Man, X. Y., Zhang, T. Y., Wang, C., & Ma, R. (2018). 我国中小微企业信用风险因素识别及测度研究 [Research on the identification and measurement of credit risk factors of small and medium-sized enterprises in China]. *journal of Central University of Finance & Economics*, (9), 46-58.
- Meng, J. (2014). 随机森林模型在财务失败预警中的应用 [Application of random forest model in financial failure early warning]. *Statistics & Decision*, 4(3), 179-181.
- Meng, J., Li, T., & Yuan, Z. (2018). 基于 ODR-BADASYN-SVM 的中小企业信用风险评估 [Credit risk assessment of SMES based on ODR - BADASYN - SVM]. *Journal of Financial Development Research*, (1), 24-31.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29(2), 449-470.
- Merton, R. C., & Bodie, Z. (1992). On the management of financial guarantees. *Financial Management*, 21(4), 87-109.
- Michels, J. (2012). Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *The Accounting Review*, 87(4), 1385-1413.
- Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28(4), 603-614.
- Modigliani, F., & Miller, M. H. (1959). The cost of capital, corporation finance, and the theory of investment: Reply. *The American Economic Review*, 49(4), 655-669.
- Mossman, C. E., Bell, G. G., Swartz, L. M., & Turtle, H. (1998). An Empirical Comparison of Bankruptcy Models. *The Financial Review*, 33(2), 33-54.
- Odom, M. D., & Sharda, R. (1990, January 15 - 19). *A neural network model for bankruptcy prediction*. International Joint Conference on Neural Networks, San Diego, CA, USA.
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109-131.

- Ooghe, H., Joos, P., & De Bourdeaudhuij, C. (1995). Financial Distress Models in Belgium: The Results of a Decade of Empirical Research. *International Journal of Accounting*, 30(3), 245-274.
- Padilla, A. J., & Pagano, M. (1997). Endogenous communication among lenders and entrepreneurial incentives. *The Review of Financial Studies*, 10(1), 205-236.
- Padilla, A. J., & Pagano, M. (2000). Sharing default information as a borrower discipline device. *European Economic Review*, 44(10), 1951-1980.
- Pagano, M., & Jappelli, T. (1993). Information sharing in credit markets. *Journal of Finance*, 48(5), 1693-1718.
- Pederzoli, C., Thoma, G., & Torricelli, C. (2013). Modelling credit risk for innovative SMEs: The role of innovation measures. *Journal of Financial Services Research*, 44(1), 111-129.
- Pederzoli, C., & Torricelli, C. (2010). A parsimonious default prediction model for Italian SMEs. *Banks and Bank System*, 5(4), 28-32.
- Pei, P., & Caoi, Y. (2016). 群组制度对 P2P 网贷平台借款成功率和借款利率的影响——基于 Prosper.com 样本数据的实证检验 [The influence of group system on the loan success rate and loan interest rate of P2P online lending platform——An empirical test based on the sample data of Prosper.com] *Economic Theory and Business Management*, (10), 5-15.
- Peng, H. F., & Yang, L. M. (2017). 抱团是否可以取暖?——群组制度对 P2P 借贷行为的影响研究 [Can holding a group keep warm?——A study on the influence of group system on P2P lending behavior]. *Foreign Economies and Management*, 39(5), 85-99.
- Petersen, M. A., & Rajan, R. G. (1994). The benefits of lending relationships: Evidence from small business data. *The Journal of Finance*, 49(1), 3-37.
- Pettway, R. H., & Sinkey, J. F. (1980). Establishing on - site bank examination priorities: an early - warning system using accounting and market information. *Journal of Finance*, 35(1), 137-150.
- Pinches, G. E., & Trieschmann, J. S. (1974). The efficiency of alternative models for solvency surveillance in the insurance industry. *Journal of Risk and Insurance*, 41(4), 563-577.
- Platt, H. D., & Platt, M. B. (1990). Development of a Class of Stable Predictive Variables: The Case of Bankruptcy Prediction. *Journal of Business Finance and Accounting*, 17(1), 31-51.
- Pope, D. G., & Sydnor, J. R. J. J. o. H. r. (2011). What' s in a Picture? Evidence of Discrimination from Prosper. com. *Journal of Human resources*, 46(1), 53-92.
- Prinzie, A., & Van den Poel, D. (2008). Random forests for multiclass classification: random multinomial Logit. *Expert systems with Applications*, 34(3), 1721-1732.
- Qian, H., Mei, Q., & Wen, X. (2013). 小微企业信贷风险评估实证研究 [Empirical study on credit risk assessment of micro and small enterprises]. *Science and Technology Management Research*, (4), 220-223.
- Qin, J., & He, Z. S. (2005, August 18-21). *A SVM face recognition method based on Gabor - featured key points*. International Conference on Machine Learning and Cybernetics, Guangzhou, China.
- Qinlan, Z., & Izumida, Y. (2013). Determinants of repayment performance of group lending in China: Evidence from rural credit cooperatives' program in Guizhou province. *China Agricultural Economic Review*, 5(3), 328-341.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.
- Quinlan, J. R. (1993, June 27-29). *Combining instance-based and model-based learning*. International Conference on Machine Learning, Amherst, MA, USA.
- Rao, C., Liu, M., Goh, M., & Wen, J. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to “Three Rurals” borrowers. *Applied Soft Computing*, 95, 106570.
- Rujoub, M. A., ook, D., & Hay, L. (1995). Using Cash Flow Ratios to Predict Business Failures.

- Journal of Managerial Issue*, 7(1), 75-90.
- Salchenberger, L. M., Cinar, E. M., & Lash, N. A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Science*, 23(4), 899-916.
- Santomero, A. M., & Vinso, J. D. (1977). Estimating the probability of failure for commercial banks and the banking system. *Journal of Banking & Finance*, 1(2), 185-205.
- Schipper, K. (1977). Financial distress in private colleges. *Journal of Accounting Research*, (15), 1-45.
- Scholes, M., & Black, F. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637-654.
- Scott, J. A. (2004). Small business and the value of community financial institutions. *Journal of Financial Services Research*, 25(2), 207-230.
- Sharda, R., & Wilson, R. L. (1996). Neural Network experiments in business – failure forecasting: Predictive performance measurement issues. *International Journal of Computational Intelligence and Organization*, 1(2), 107-117.
- Sheng, X., Li, B., & Zhang, D. (2016). 基于数据挖掘的上市公司信用评级变动预测 [Forecast of credit rating changes of listed companies based on data mining]. *Statistics & Decision*, 15(3), 159-162.
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355-374.
- Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213-225.
- Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, 71(3), 393-410.
- Sun, A., Lim, E. P., & Ng, W. K. (2002, December 28-31). *Web classification using support vector machine*. International Workshop on Web Information and Data Management, New York, USA.
- Sun, Y., Chai, N., Dong, Y., & Shi, B. (2022). Assessing and predicting small industrial enterprises' credit ratings: A fuzzy decision-making approach. *International Journal of Forecasting*, 38(3), 1158-1172.
- Tanak, K., Kinkyō, T., & Hamori, S. (2016). Random forests – based early warning system for bank failures. *Economics Letters*, 148, 118-121.
- Tian, J., Gu, H., & Liu, W. (2011). Imbalanced classification using support vector machine ensemble. *Neural Computing & Applications*, 20(2), 203-209.
- Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science*, 174, 150-160.
- Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics Applied Probability*, 57, 1-436.
- Tra Pham, T. T., & Lensink, R. (2008). Household borrowing in Vietnam: a comparative study of default risks of formal, informal and semi – formal credit. *Journal of Emerging Market Finance*, 7(3), 237-261.
- Trippi, R., & Turban, E. (1996). *Neural Networks in finance and investing*. Irwin Professional Publishing.
- Tsai, C. F., & Wu, J. W. (2008). Using Neural Network Ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4), 2639-2649.
- Umezawa, Y., & Mori, H. (2008). Credit risk evaluation of power market players with random forest. *IEEE Transactions on Power and Energy*, 128(1), 165-172.
- Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons. Inc.
- Wang, C., Wan, H., & Zhang, W. (1999). 基于神经网络技术的商业银行信用风险评估 [Commercial bank credit risk assessment based on neural network technology]. *Systems Engineering Theory & Practice*, 19(9), 24-32.

- Wang, C., Zhang, W., Zhao, X., & Wang, J. (2019). Soft information in online peer-to-peer lending: Evidence from a leading platform in China. *Electronic Commerce Research and Applications*, 36, 100873.
- Wang, H. J., & He, L. (2015). 借款描述对 P2P 网络借贷行为影响的实证研究 [An empirical study on the influence of loan description on P2P network lending behavior]. *Financial Economics Research*, 30(1), 77-85.
- Wang, J. M. (2004). 我国基础信用信息共享机制的问题及对策 [Problems and countermeasures of China's basic credit information sharing mechanism]. *Journal of Intelligence*, (5), 23-24, 27.
- Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science*, 199, 1128-1135.
- Wang, T., Liu, R., & Qi, G. (2022). Multi-classification assessment of bank personal credit risk based on multi-source information fusion. *Expert Systems with Applications*, 191, 116236.
- Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning. *Procedia Computer Science*, 174, 141-149.
- Wang, Y. J., & Sheng, D. (2012). 政治关联与企业的契约实施环境 [Political connection and the contract implementation environment of enterprises] *China Economic Quarterly*, 11(4), 1193-1218.
- Wen, X., & Han, X. (2017). 我国科技型中小企业信用风险评价模型——基于 MLP 神经网络的实证分析 [Credit Risk evaluation model of China's high - tech SMEs: an empirical analysis based on MLP neural network]. *Science Technology and Industry*, 17(12), 159-166.
- Wilcox, J. W. (1971). A Simple Theory of Financial Ratios as Predictors of Failures. *Journal of Accounting Research*, (8), 389-395.
- Wilcox, J. W. (1973). A prediction of business failure using accounting data. *Journal of Accounting Research*, 11, 163-179.
- Williamson, S. D. (1987). Costly monitoring, loan contracts, and equilibrium credit rationing. *The Quarterly Journal of Economics*, 102(1), 135-145.
- X, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications*, 36(2), 2625-2632.
- Xia, H. (2015). 基于主成分分析和支持向量回归机组合模型的电子商务信用风险度预测研究 [E - commerce credit risk prediction based on PCA - SVR]. *Journal of Modern Information*, (1), 76-79.
- 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 Credit Risks Evaluation Model Based on Support - vector Machine Regression Ensemble]. *Credit Reference*, (4), 21-27.
- Xiao, B., Bai, W., Yao, Y., & Li, X. (2016). 基于 LS-SVM 的小微企业信用评估研究 [Micro and small businesses credit evaluation research based on LSSVM]. *Journal of Audit & Economics*, (6), 102-112.
- Xiong, X., Ma, J., & Zhao, W. (2009). 供应链金融的信用风险分析 [Credit risk analysis of supply chain finance]. *Nankai Business Review*, (12), 92-98.
- Xu, X., Gao, Z., & Zhao, N. (2014). 基于支持向量机的小企业信用评级最适样本量的研究 [Study on the optimal sample size of small business credit rating based on SVM]. *Journal of Huaiyin Teachers College (Natural Science Edition)*, 13(1), 40-43.
- Xu, Y., & Pan, M. (2016). 层次分析法和支持向量机在个人信用评估中的应用 [Application of analytic process and support vector machine in personal credit evaluation]. *Management Science in China*, 24(1), 106-112.
- Yang, K. (2015). 修正 KMV 模型在创业板上市公司信用风险度量中的应用分析

- [Analysis on the application of modified KMV model in credit risk measurement of listed companies on the growth enterprise market]. *West China Finance*, 17(1), 21-24.
- Yang, K., Yuan, H., & Lau, R. Y. K. (2022). PsyCredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. *Expert Systems with Applications*, 198, 116847.
- Yao, B. (2016). 信用评价是否抑制了网贷行为中的地区与学历偏好 [Does credit evaluation inhibit regional and academic preference in online lending behavior?]. *Finance & Trade Economics*, (7), 147-161.
- Yao, G., Hu, X., & Wang, G. (2022). A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. *Expert Systems with Applications*, 200, 117002.
- Ye, D., & Chen, X. (2017). 标点与字数会影响网络借贷吗——来自人人贷的经验证据 [Do Punctuation and Word Count Affect Online Lending: Empirical Evidence from Renrendai] *Finance & Trade Economics*, 38(05), 65-79.
- Yeh, C. C., Chi, D. J., & Lin, Y. R. (2014). Going – concern prediction using hybrid random forests and rough set approach. *Information Sciences*, 254, 98-110.
- 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(3), 166-172.
- Yin, C., Jiang, C., Jain, H. K., & Wang, Z. (2020). Evaluating the credit risk of SMEs using legal judgments. *Decision Support Systems*, 136, 113364.
- Yu, B., Li, C., Mirza, N., & Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, 174, 121255.
- Yu, L., & Zhang, X. (2021). Can small sample dataset be used for efficient internet loan credit risk assessment? Evidence from online peer to peer lending. *Finance Research Letters*, 38, 101521.
- Yu, L., Zhang, X., & Yin, H. (2022). An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity. *Expert Systems with Applications*, 202, 117363.
- Zai, D., Zhang, J., & Cao, Y. (2007). KMV 模型在上市公司信用风险管理中的应用 [Application of KMV Model in Credit Risk Management of Listed Companies]. *Industrial Technology & Economy*, (1), 126-127.
- Zeller, M. (1998). Determinants of repayment performance in credit groups: The role of program design, intragroup risk pooling, and social cohesion. *Economic Development and Cultural Change*, 46(3), 599-620.
- Zeng, H., Wang, Z., & Cui, X. (2013). 基于 SVM 的中小企业集合债券融资个体信用风险度量研究 [Research on individual credit risk measurement of SME collective bond financing based on SVM]. *Journal of Central South University (Social Science)*, 19(2), 8-11, 19.
- Zeng, S., & Wang, F. (2013). 基于 KMV 模型的制造业上市公司信用风险评价研究 [The empirical research on credit risk of listed manufacturing companies from the view of KMV model]. *Forecasting*, 32(2), 60-63, 69.
- Zhang, H., & Li, H. (2017). BP 神经网络模在中小企业融资征信评估中的应用 [Application of BP neural network model in the evaluation of SMEs' financing credit]. *Modernization of Shopping Malls*, (18), 172-173.
- Zhang, H., Shi, Y., Yang, X., & Zhou, R. (2021). A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance. *Research in International Business and Finance*, 58, 101482.
- Zhang, H. Y., & Jiang, H. B. (2008). 我国银行信用信息共享内生性分析 [An analysis of the

- endogeneity of China's inter-bank credit information sharing]. *Shanghai Economic Research*, (10), 99-104.
- Zhang, L., Yang, Z., & Chen, S. (2004). KMV 模型在上市公司信用风险评价中的应用研究 [Application Research of KMV Model in Credit Risk Assessment of Listed Companies]. *Systems Engineering - Theory & Practice*, 22(11), 84-89.
- Zhang, N., & Zhang, J. (2010). 改进的 KMV 模型在我国上市公司信用风险度量中的应用 [Application of improved KMV model in credit risk measurement of listed companies in China]. *Forecasting*, 29(5), 48-52.
- Zhang, Q., Hu, L., & Wang, Y. (2015). Research on bank credit risk evaluation model based on Logit and SVM. *System Engineering Theory and Practice*, 35(7), 1784-1790.
- Zhang, R., & Liang, Q. (2006). 我国银行间信用信息共享行为与征信体系构建分析 [Analysis on China's inter-bank credit information sharing behavior and credit information system construction] *Business Economics and Management*, (12), 37-41, 70.
- Zhang, X., Yu, L., Yin, H., & Lai, K. K. (2022). Integrating data augmentation and hybrid feature selection for small sample credit risk assessment with high dimensionality. *Computers & Operations Research*, 146, 105937.
- Zhao, F., Lu, Y., Li, X., Wang, L., Song, Y., Fan, D., Zhang, C., & Chen, X. (2022). Multiple imputation method of missing credit risk assessment data based on generative adversarial networks. *Applied Soft Computing*, 126, 109273.
- Zhao, Y. (2013). 我国小微企业信用评级指标体系研究 [Research on credit rating index system of China's micro and small enterprise] [Master's thesis], Jilin University of Finance and Economics.
- Zhao, Y. (2015). 随机森林算法在地方融资平台主体信用评级中的应用 [Application of random forest algorithm in subject credit rating of local financing platform]. *Finance and Accounting Monthly*, 35(7), 1784-1790.
- Zhou, B. (2021). 小微企业互联网融资信用评价研究 [Research on credit evaluation of small and micro enterprises' internet financing] [Doctoral dissertation], Beijing Jiaotong University.
- Zhou, J. (2009). 修正的 KMV 模型在上市公司信用风险度量中的应用分析 [Application analysis of modified KMV model in credit risk measurement of listed companies]. *Communication of Finance and Accounting*, (15), 8-10.
- Zhou, Z. (2016). 机器学习 [Machine learning]. Tsinghua University Press.
- Zhu, Y., Xie, C., Sun, B., Wang, G. J., & Yan, X. G. (2016). Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models. *Sustainability*, 8(5), 433.
- Zhuang, L., Zhou, Q., & Fu, Y. (2015). P2P 网络借贷用途、投资偏向效应与软信息价值 [P2P network lending use, investment bias effect and soft information value]. *International Business (Journal of University of International Business and Economics)*, (6), 77-85.
- Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, (22), 59-82.

Webliography

- Agrawal, A. K., Catalini, C., & Goldfarb, A. (2011, February 01). *The geography of crowdfunding*. National Bureau of Economic Research. Retrieved October 20, 2021, from https://www.nber.org/system/files/working_papers/w16820/w16820.pdf
- Bernanke, B. S., & Blinder, A. S. (1988, March 01). *Credit, money, and aggregate demand*. National Bureau of Economic Research. Retrieved October 20, 2021, from https://www.nber.org/system/files/working_papers/w2534/w2534.pdf
- Chen, J., Liu, B., & Huang, D. (2016, May 08). *Protein remote homology detection based on an ensemble learning approach*. BioMed Research International. Retrieved October 20, 2021, from <https://www.hindawi.com/journals/bmri/2016/5813645/>
- Dwyer, D., Kocagil, A., Stein, R., & Wingate, A. (2004, January 01). *Moody ' s KMV risk scale v3.1 United States: modeling methodology*. Moody. Retrieved October 20, 2021, from https://www.researchgate.net/publication/215991512_THE_MOODY'S_KMV_EDF_RISK_CALC_V31_MODEL
- Falavigna, G. (2006, July 01). *Models for default risk analysis focus on artificial neural networks*. Institute for Economic Research on Firms and Growth. Retrieved October 20, 2021, from http://www.ceris.cnr.it/ceris/workingpaper/2006/WP_10_06_FALAVIGNA.pdf
- Fama, E. F., Jensen, M. C., Long, J. B., Ruback, R. S., Schwert, G. W., Smith, C. W., & Warner, J. B. (2003, November 28). *Clinical papers and their role in the development of financial economics*. Social Science Research Network. Retrieved October 20, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=466661
- Freedman, S., & Jin, G. Z. (2008, November 14). *Do social networks solve information problems for peer-to-peer lending? Evidence from Prosper. com*. Social Science Research Network. Retrieved October 21, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1304138
- Kira, K., & Rendell, L. A. (2014, June 27). *A practical approach to feature selection*. Elsevier. Retrieved October 21, 2021, from <https://www.sciencedirect.com/science/article/pii/B9781558602472500371>
- Kononenko, I. (1994, January 01). *Estimating attributes: Analysis and extensions of RELIEF*. Springer. Retrieved January 10, 2022, from https://link.springer.com/content/pdf/10.1007/3-540-57868-4_57.pdf
- Nyfeler, M. (2000, November 23). *Modeling dependencies in credit risk management*. Risk Lab. Retrieved October 20, 2021, from <http://www2.risklab.ch/ftp/papers/DiplomaThesisNyfeler.pdf>
- Ravina, E. (2019, March 07). *Love & loans: The effect of beauty and personal characteristics in credit markets*. Social Science Research Network. Retrieved October 20, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1101647

[This page is deliberately left blank.]

Annex A: Tables

Table A1 Classification standard of tax credit rating

Rating	Score
A	[90,100]
B	[70,90)
M	[70,100], and 1. Newly established enterprises. 2. Evaluated enterprises that have no production and operation business income and have an annual evaluation score of 70 or more.
C	[40,70)
D	[0,40)

Source: “Announcement on the Matters Related to Tax Credit Evaluation” issued by the State Administration of Taxation.

Table A2 Confusion matrix of classification results

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	TP	FN
-1(negative)	FP	TN

Source: Author.

Table A3 Kernel functions commonly used in SVM

Name	Expression	Parameter
Linear kernel	$K(x_i, x_j) = x_i^T x_j$	None
Polynomial kernel	$K(x_i, x_j) = (x_i^T x_j)^d$	$d \geq 1$ is the degree of the polynomial
Radial Basis Function kernel	$K(x_i, x_j) = \exp(-\frac{\ x_i - x_j\ _2^2}{2\sigma^2})$	$\sigma > 0$ is the bandwidth
Laplace kernel	$K(x_i, x_j) = \exp(-\frac{\ x_i - x_j\ _2}{\sigma})$	$\sigma > 0$
Sigmoid kernel	$K(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$	$\beta > 0, \theta < 0$

Source: Author.

Table A4 Distribution of tax credit rating of different samples

Tax credit rating	Non-passing	Defaulting	Non-defaulting
A	210	17	256
B	899	34	290
C	114	3	0
D	81	0	0

Source: Author.

Table A5 Basic features of defaulting sample and non-defaulting sample

Features	Non-defaulting	Defaulting	Difference	P value
Loan application amount (ten thousands)	156.52	67.09	89.43	0.000***
Number of employees	15.84	12.37	3.47	0.284
Registered capital (ten thousands)	34.20	50.99	-16.79	0.106
Enterprise's duration	9.03	8.06	0.97	0.083*

Note: ***, **, and * represent significant at the level of 1%, 5%, and 10%, respectively.

Source: Author.

Table A6 Default rate of different business registration type

Registration type	Number of defaulting enterprises	Number of enterprises (defaulting + non-defaulting)	Default rate (%)
Private limited company	41	363	11.29
Other limited company	13	231	5.63
Sole proprietorship	0	6	0

Source: Author.

Table A7 Default rate of different industries

Industry	Number of defaulting enterprises	Number of enterprises (defaulting + non-defaulting)	Default rate (%)
Other industries	9	58	15.52
Wholesale and retail	16	190	8.42
Manufacturing	29	352	8.24

Source: Author.

Table A8 Default rate of different tax credit ratings

Tax credit rating	Number of defaulting enterprises	Number of enterprises (defaulting + non-defaulting)	Default rate (%)
B	37	327	11.31
A	17	273	6.23

Source: Author.

Table A9 Default rate of different taxpayer types

Taxpayer type	Number of defaulting enterprises	Number of enterprises (defaulting + non-defaulting)	Default rate (%)
Small-scale taxpayer	1	10	10
General taxpayer	53	590	8.98

Source: Author.

Table A10 Distribution of the number of rules triggered

Number of rules triggered	1	2	3	4	5	6
Number of enterprises	676	316	157	83	38	16
Percentage	51.92%	24.27%	12.06%	6.37%	2.92%	1.23%
Number of rules triggered	7	8	9	10	11	12
Number of enterprises	5	6	1	2	1	1
Percentage	0.38%	0.46%	0.08%	0.15%	0.08%	0.08%

Source: Nantong Branch of Jiangsu Bank.

Table A11 Sub-samples generated from different feature sets

Sample number	Features contained
Sample 1	Basic features
Sample 2	Financial features
Sample 3	Tax features
Sample 4	Basic features + Financial features
Sample 5	Basic features + Tax features
Sample 6	Financial features + Tax features
Sample 7	Basic features + Financial features + Tax features

Source: Author.

Table A12: Optimal parameter values of SVM classifier on each sample

Sample number	C	σ
Sample 1	7.38105263	0.00088587
Sample 2	7.31578947	0.00088587
Sample 3	8.73684211	0.04832930
Sample 4	0.90000000	0.00335982
Sample 5	7.52631579	0.18329807
Sample 6	7.94736842	0.00088587
Sample 7	3.84210526	0.01274275

Source: Author.

Table A13 Number of enterprises in training and testing sets

Label	Training set	Testing set
-1(negative)	385	161
+1(positive)	35	19

Source: Author.

Table A14 Confusion matrix of SVM classifier on sample 1

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	19 (TP)	0 (FN)
-1(negative)	63 (FP)	98 (TN)

Source: Author.

Table A15 Performance metrics of SVM classifier on sample 1

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
96.50%	65.00%	100.00%	23.17%	37.62%	54.44%	0.00%

Source: Author.

Table A16 Confusion matrix of optimal threshold SVM classifier on sample 1

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	12 (FP)	149 (TN)

Source: Author.

Table A17 Performance metrics of optimal threshold SVM classifier on sample 1

Accuracy	Recall	Precision	F1	Approval rate	Default rate
92.22%	89.47%	58.62%	70.83%	83.89%	1.32%

Source: Author.

Table A18 Confusion matrix of SVM classifier on sample 2

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	13 (TP)	6 (FN)
-1(negative)	10 (FP)	151(TN)

Source: Author.

Table A19 Performance metrics of SVM classifier on sample 2

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
86.56%	91.11%	68.42%	56.52%	61.90%	87.22%	3.82%

Source: Author.

Table A20 Confusion matrix of optimal threshold SVM classifier on sample 2

True labels	Predicated labels	
	+1(positive)	-1(negative)

+1(positive)	15 (TP)	4 (FN)
-1(negative)	16 (FP)	145 (TN)

Source: Author.

Table A21 Performance metrics of optimal threshold SVM classifier on sample 2

Accuracy	Recall	Precision	F1	Approval rate	Default rate
88.89%	78.95%	48.39%	60.00%	82.78%	2.68%

Source: Author.

Table A22 Confusion matrix of SVM classifier on sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	3 (FP)	158 (TN)

Source: Author.

Table A23 Performance metrics of SVM classifier on sample 3

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
97.03%	97.22%	89.47%	85.00%	87.18%	88.89%	1.25%

Source: Author.

Table A24 Confusion matrix of optimal threshold SVM classifier on sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	2 (FP)	159 (TN)

Source: Author.

Table A25 Performance metrics of optimal threshold SVM classifier on sample 3

Accuracy	Recall	Precision	F1	Approval rate	Default rate
97.78%	89.47%	89.47%	89.47%	89.44%	1.24%

Source: Author.

Table A26 Confusion matrix of SVM classifier on sample 4

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	13 (TP)	6 (FN)
-1(negative)	5 (FP)	156 (TN)

Source: Author.

Table A27 Performance metrics of SVM classifier on sample 4

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
93.40%	93.89%	68.42%	72.22%	70.27%	90.00%	3.70%

Source: Author.

Table A28 Confusion matrix of optimal threshold SVM classifier on sample 4

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	15 (TP)	4 (FN)
-1(negative)	6 (FP)	155(TN)

Source: Author.

Table A29 Performance metrics of optimal threshold SVM classifier on sample 4

Accuracy	Recall	Precision	F1	Approval rate	Default rate
94.44%	78.95%	71.43%	75.00%	88.33%	2.52%

Source: Author.

Table A30 Confusion matrix of SVM classifier on sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	15 (TP)	4(FN)
-1(negative)	1 (FP)	160(TN)

Source: Author.

Table A31 Performance metrics of SVM classifier on sample 5

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
94.21%	97.22%	78.95%	93.75%	85.71%	91.11%	2.44%

Source: Author.

Table A32 Confusion matrix of optimal threshold SVM classifier on sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2(FN)
-1(negative)	2 (FP)	159(TN)

Source: Author.

Table A33 Performance metrics of optimal threshold SVM classifier on sample 5

Accuracy	Recall	Precision	F1	Approval rate	Default rate
97.78%	89.47%	89.47%	89.47%	89.44%	1.24%

Source: Author.

Table A34 Confusion matrix of SVM classifier on sample 6

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	16 (TP)	3(FN)
-1(negative)	3 (FP)	158(TN)

Source: Author.

Table A35 Performance metrics of SVM classifier on sample 6

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
93.59%	96.67%	84.21%	84.21%	84.21%	89.44%	1.86%

Source: Author.

Table A36 Confusion matrix of optimal threshold SVM classifier on sample 6

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2(FN)
-1(negative)	3 (FP)	158(TN)

Source: Author.

Table A37 Performance metrics of optimal threshold SVM classifier on sample 6

Accuracy	Recall	Precision	F1	Approval rate	Default rate
97.22%	89.47%	85.00%	87.18%	88.89%	1.25%

Source: Author.

Table A38 Confusion matrix of SVM classifier on sample 7

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	12 (TP)	7(FN)
-1(negative)	1 (FP)	160(TN)

Source: Author.

Table A39 Performance metrics of SVM classifier on sample 7

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
95.85%	95.56%	63.16%	92.31%	75.00%	92.78%	4.19%

Source: Author.

Table A40 Confusion matrix of optimal threshold SVM classifier on sample 7

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	16 (FP)	145 (TN)

Source: Author.

Table A41 Performance metrics of optimal threshold SVM classifier on sample 7

Accuracy	Recall	Precision	F1	Approval rate	Default rate
90.00%	89.47%	51.52%	65.38%	81.67%	1.36%

Source: Author.

Table A42 Parameter optimization results of grid method for SVM

Sample	C	σ
New Sample3	8.73684	0.04833
New Sample5	7.526315789	0.183298071

Source: Author.

Table A43 Confusion matrix of SVM classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	5 (FP)	156(TN)

Source: Author.

Table A44 Performance metrics of SVM classifier on new sample 3

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
96.68%	96.11%	89.47%	75.00%	83.72%	86.74%	1.27%

Source: Author.

Table A45 Confusion matrix of optimal threshold SVM classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	18 (TP)	1 (FN)
-1(negative)	4 (FP)	157 (TN)

Source: Author.

Table A46 Performance metrics of optimal threshold SVM classifier on new sample 3

Accuracy	Recall	Precision	F1	Approval rate	Default rate
97.22%	94.74%	81.82%	87.80%	87.78%	0.63%

Source: Author.

Table A47 Confusion matrix of SVM classifier on new sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	1 (FP)	160 (TN)

Source: Author.

Table A48 Performance metrics of SVM classifier on new sample 5

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
94.31%	98.33%	89.47%	94.44%	91.89%	90.00%	1.23%

Source: Author.

Table A49 Confusion matrix of optimal threshold SVM classifier on new sample5

True labels	Predicated labels	
	+1(positive)	-1(negative)

	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	0 (FP)	161 (TN)

Source: Author.

Table A50 Performance metrics of optimal threshold SVM classifier on new sample 5

Accuracy	Recall	Precision	F1	Approval rate	Default rate
98.89%	89.47%	100.00%	94.44%	90.56%	1.23%

Source: Author.

Table A51 Parameter tuning results of the LR algorithm

Sample	C	mean_auc
New sample 3	4.63158	0.90164
New sample 5	14.10526	0.89725

Source: Author.

Table A52 Confusion matrix of LR classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	1 (FP)	160 (TN)

Source: Author.

Table A53 Performance metrics of LR classifier on new sample 3

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
94.31%	98.33%	89.47%	94.44%	91.89%	90.00%	1.23%

Source: Author.

Table A54 Confusion matrix of optimal threshold LR classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	2 (FP)	159 (TN)

Source: Author.

Table A55 Performance metrics of optimal threshold LR classifier on new sample 3

Accuracy	Recall	Precision	F1	Approval rate	Default rate
97.78%	89.47%	89.47%	89.47%	89.44%	1.24%

Source: Author.

Table A56 Confusion matrix of LR classifier on new sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	4 (FP)	157 (TN)

Source: Author.

Table A57 Performance metrics of LR classifier on new sample 5

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
95.23%	96.67%	89.47%	80.95%	85.00%	88.33%	1.26%

Source: Author.

Table A58 Parameter tuning results of BPNN algorithm

Sample	α	hidden layer sizes	mean_auc
New sample 3	0.000001	(100,50)	0.86973
New sample 5	0.000001	(50,50)	0.86561

Source: Author.

Table A59 Confusion matrix of BPNN classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	1 (FP)	160 (TN)

Source: Author.

Table A60 Performance metrics of BPNN classifier on new sample 3

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
92.11%	98.33%	89.47%	94.44%	91.89%	90.00%	1.23%

Source: Author.

Table A61 Confusion matrix of BPNN classifier on new sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	2 (FP)	159 (TN)

Source: Author.

Table A62 Performance metrics of BPNN classifier on new sample 5

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
95.62%	97.78%	89.47%	89.47%	89.47%	89.44%	1.24%

Source: Author.

Table 63 Parameter tuning results of the RF algorithm

Sample	n-estimators	mean auc
New sample 3	19	0.890752
New sample 5	26	0.866561

Source: Author.

Table A64 Confusion matrix of RF classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	1 (FP)	160 (TN)

Source: Author.

Table A65 Performance metrics of RF classifier on new sample 3

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
96.03%	98.33%	89.47%	94.44%	91.89%	90.00%	1.23%

Source: Author.

Table A66 Confusion matrix of RF classifier on new sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	0 (FP)	161 (TN)

Source: Author.

Table A67 Performance metrics of RF classifier on new sample 5

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
96.70%	98.89%	89.47%	100.00%	94.44%	90.56%	1.23%

Source: Author.

Table A68 Confusion matrix of TSVM classifier on new sample 3

True labels	Predicated labels	
	+1(positive)	-1(negative)

+1(positive)	17 (TP)	2 (FN)
-1(negative)	1 (FP)	160 (TN)

Source: Author.

Table A69 Performance metrics of TSVM classifier on new sample 3

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
96.03%	98.33%	89.47%	94.44%	91.89%	90.00%	1.23%

Source: Author.

Table A70 Confusion matrix of TSVM classifier on new sample 5

True labels	Predicated labels	
	+1(positive)	-1(negative)
+1(positive)	17 (TP)	2 (FN)
-1(negative)	0 (FP)	161 (TN)

Source: Author.

Table A71 Performance metrics of TSVM classifier on new sample 5

AUC	Accuracy	Recall	Precision	F1	Approval rate	Default rate
96.70%	98.89%	89.47%	100.00%	94.44%	90.56%	1.23%

Source: Author.

[This page is deliberately left blank.]

Annex B: Figures

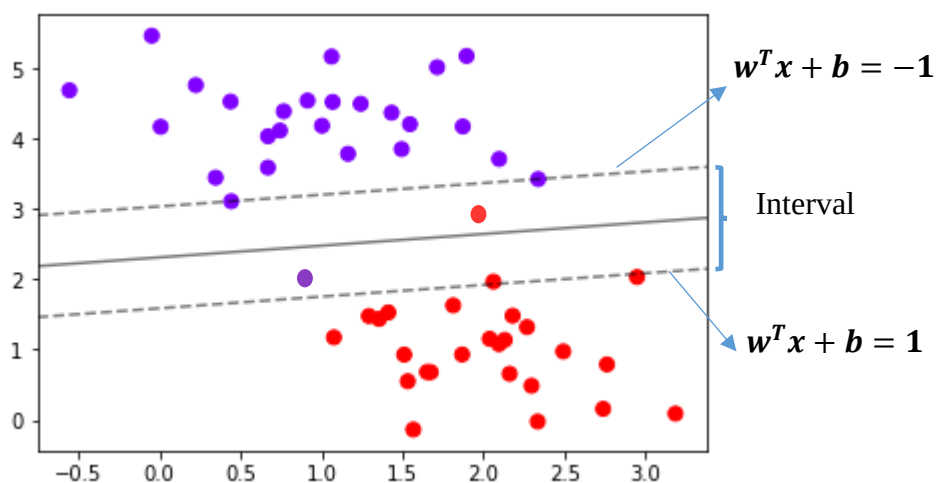


Figure B1 SVM and interval

Source: Author.

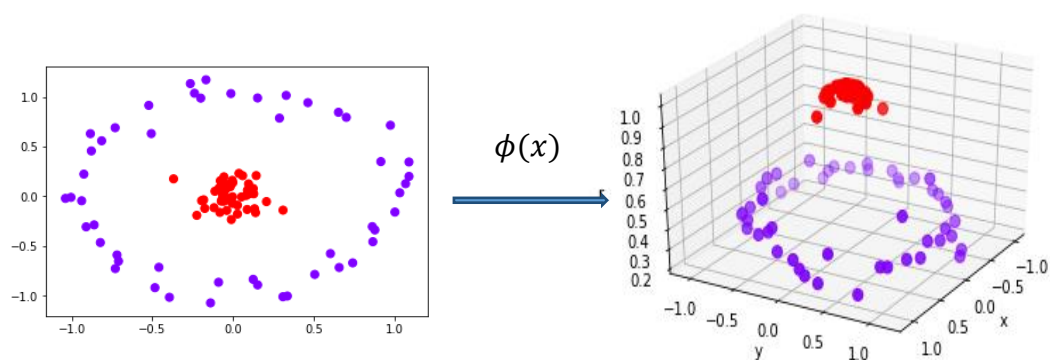
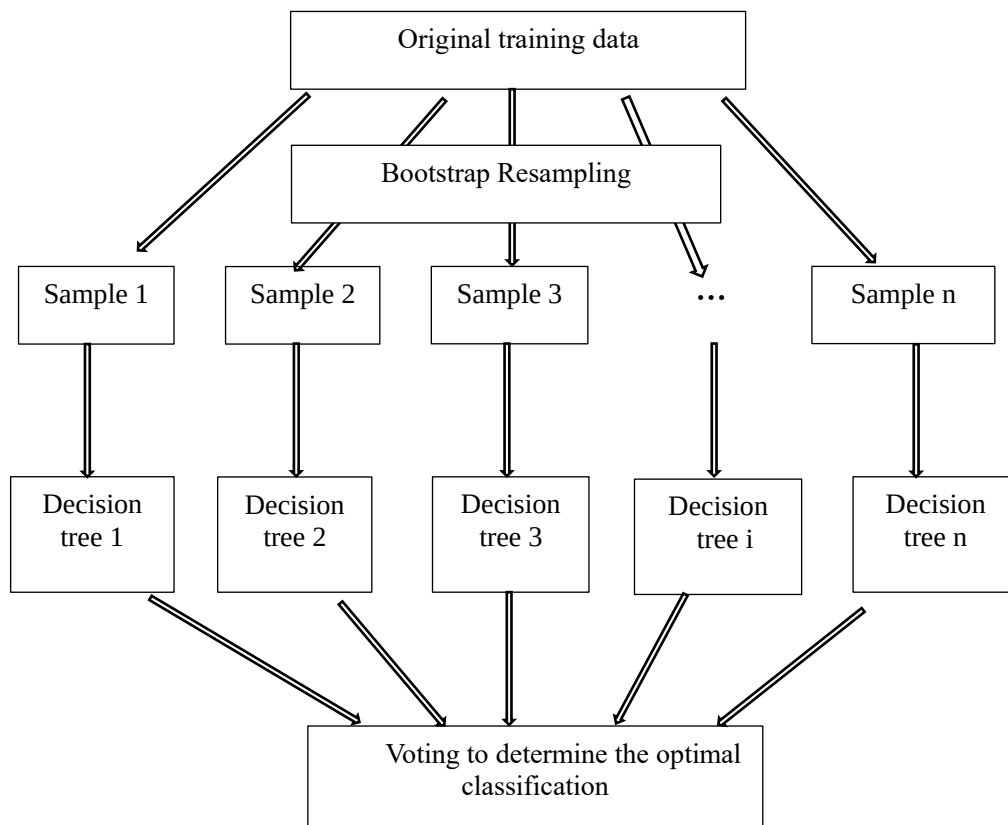


Figure B2 Mapping from low-dimensional space to high-dimensional space

Source: Author.



FigureB3 Schematic diagram of RF

Source: Author.

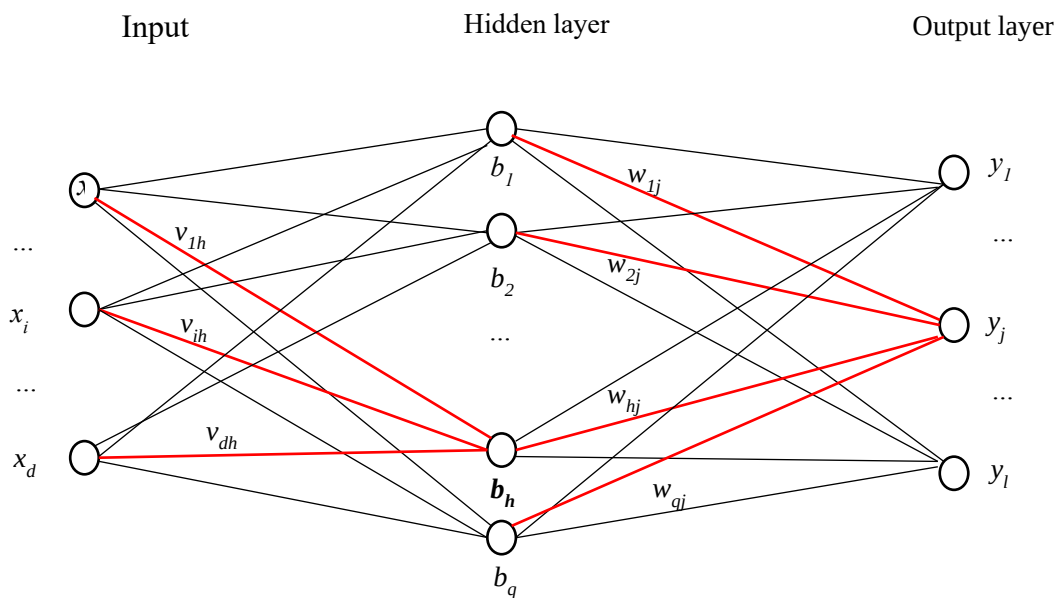


Figure B4 Schematic diagram of BPNN

Source: Author.

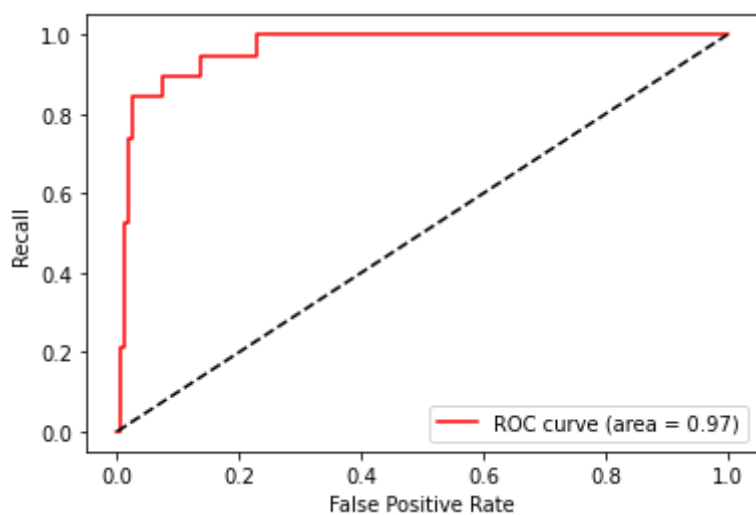


Figure B5 ROC curve of SVM classifier on sample 1

Source: Author.

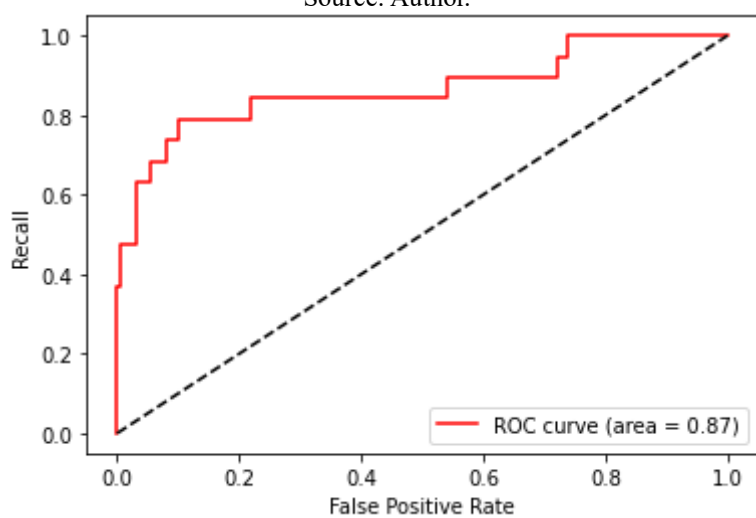


Figure B6 ROC curve of SVM classifier on sample 2

Source: Author.

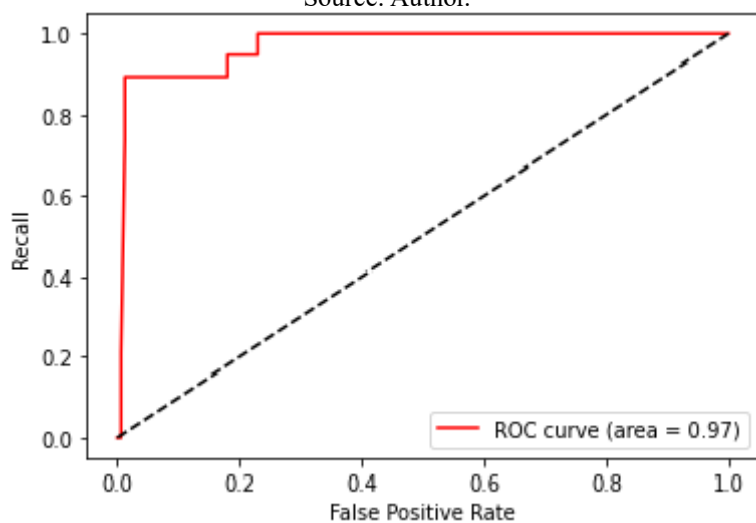


Figure B7 ROC curve of SVM classifier on sample 3

Source: Author.

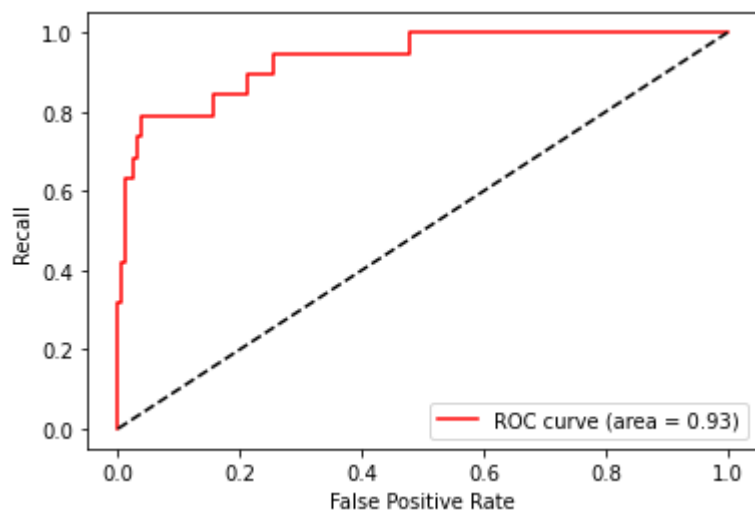


Figure B8 ROC curve of SVM classifier on sample 4

Source: Author.

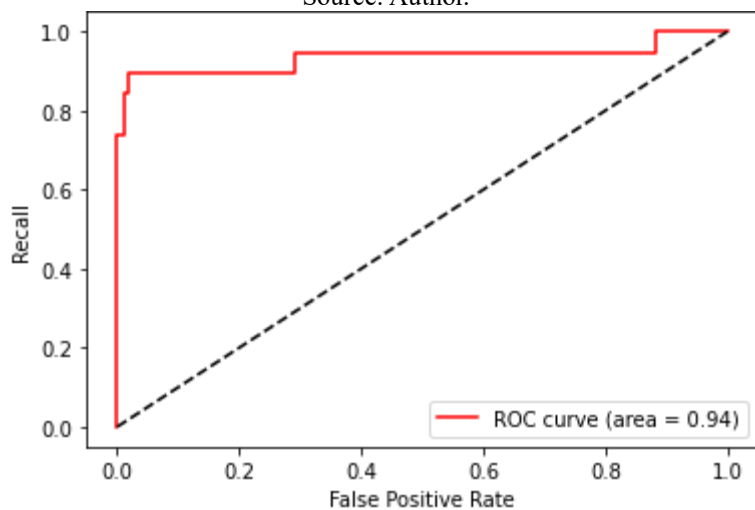


Figure B9 ROC curve of SVM classifier on sample 6

Source: Author.

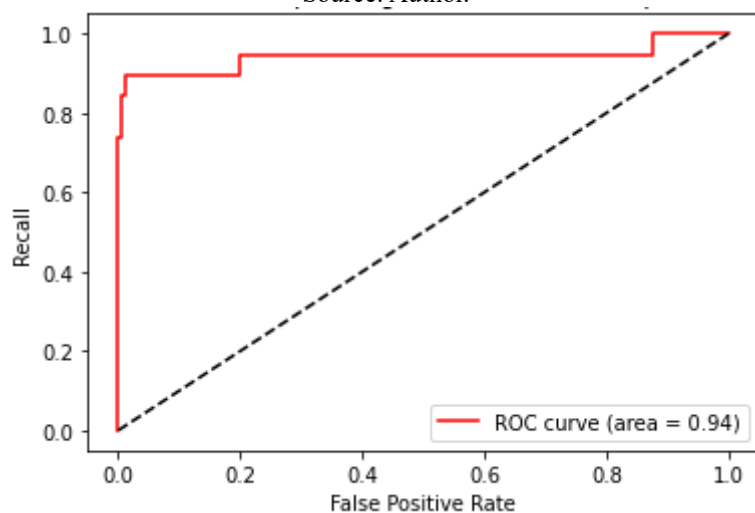


Figure B10 ROC curve of SVM classifier on sample 5

Source: Author.

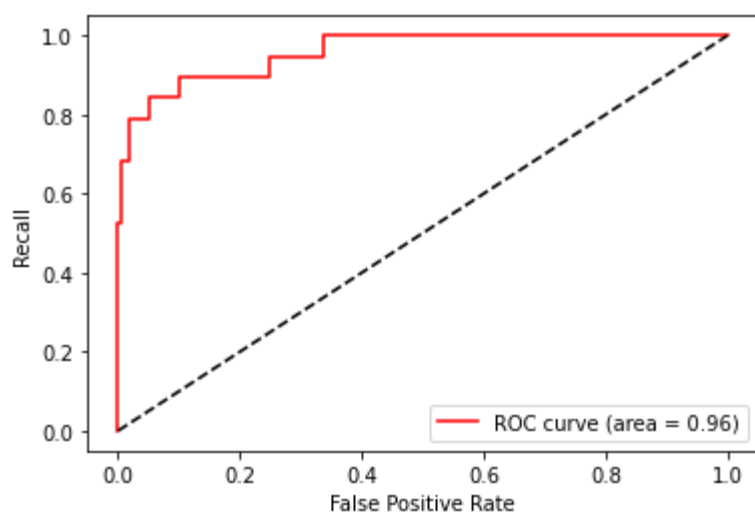


Figure B11 ROC curve of SVM classifier on sample 7

Source: Author.

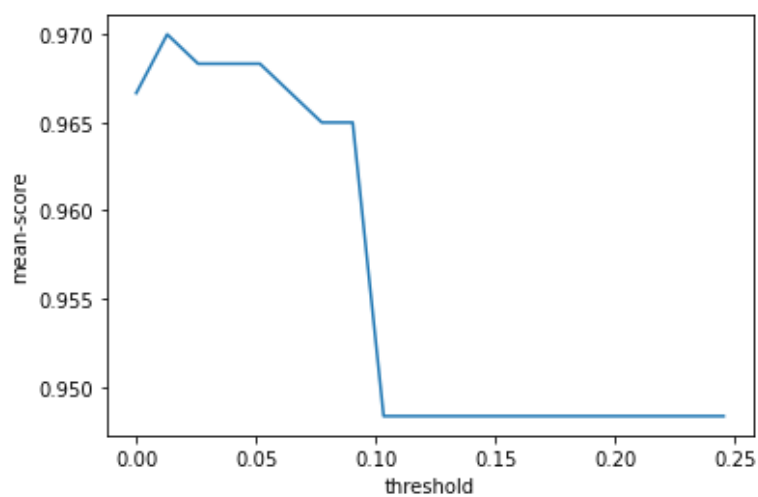


Figure B12 Importance threshold and average accuracy for sample 3

Source: Author.

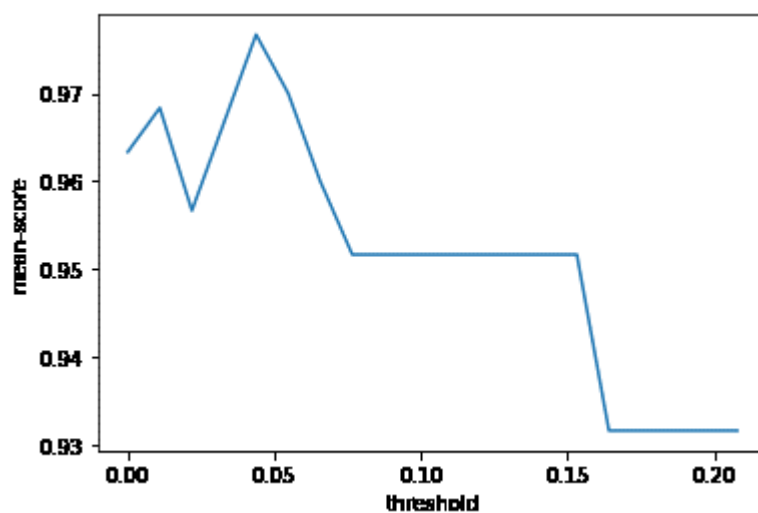


Figure B13 Importance threshold and average accuracy for sample 5

Source: Author.

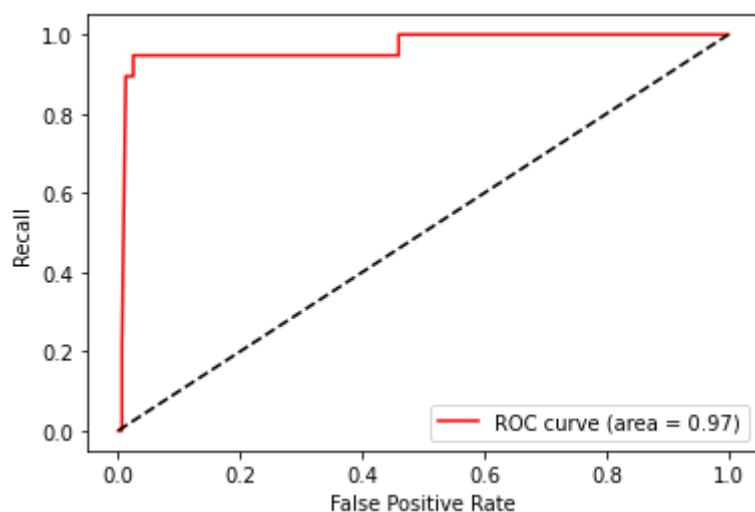


Figure B14 ROC curve of SVM classifier on new sample 3

Source: Author.

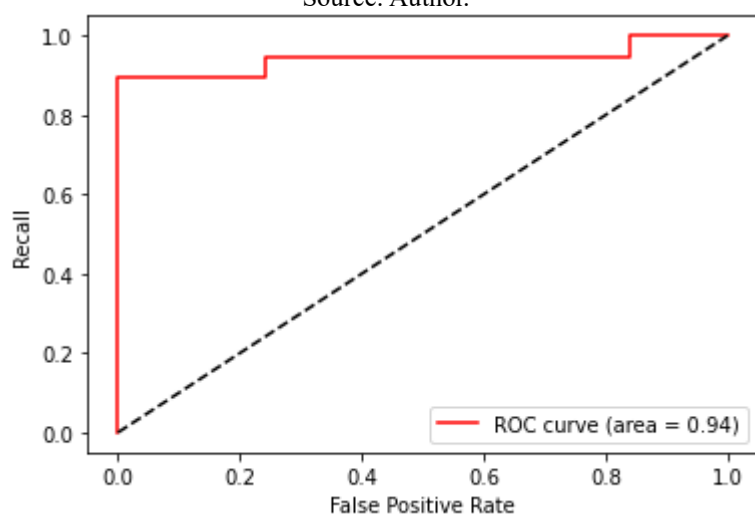


Figure B15 ROC curve of SVM classifier on new sample 5

Source: Author.

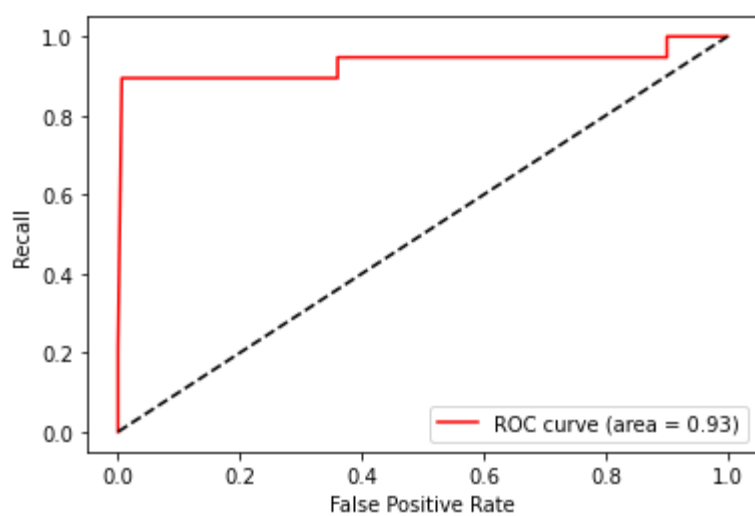


Figure B16 ROC curve of LR classifier on new sample 3

Source: Author.