Empirical Research on Credit Evaluation of Real Estate Industry Based on Multi-source Data Fusion

Jiahui Liu

Nanjing University of Science and Technology

Abstract. In this paper, based on the evaluation of corporate credit risk using traditional financial indicators, we will combine with non-financial data, define whether the listed company has ST in its abbreviation as a default fact, and establish a logistic model to quantitatively analyze the credit risk. In this paper, 123 listed companies in the real estate industry are used as samples, and the financial data, number of employees, percentage of shareholding of the largest shareholder, and CEO's education disclosed in the 2019 and 2020 annual reports are obtained through the Guotaian and CNRDS databases. Using principal component analysis to extract principal component factors from 21 financial data and combining non-financial indicators, a logistic model based on financial indicators and a logistic model based on financial and non-financial data were constructed to study the impact of non-financial data on company credit risk. It is found that non-financial indicators reflecting company size as well as corporate governance ability have a certain degree of influence on the credit risk of the company, i.e., the prediction results of the logistic model based on financial and non-financial data are more effective.

Keywords: Credit evaluation; Real estate industry; Logistic model; Non-financial data.

1. Introduction

In today's environment of deepening economic globalization, how to better manage financial risks and make financial markets stable is crucial. The main risk faced by financial institutions is credit risk, which is more difficult to predict and perceive than other financial risks. At the same time, investors, financial institutions and researchers are paying more and more attention to it. However, how to accurately assess the risk of use is still a problem worthy of study and analysis.

The real estate industry plays a very important role in China's economic structure. During the critical period when the epidemic has an impact on the economy, a considerable number of real estate enterprises are affected by the epidemic, which increases the credit risk of enterprises. And although the prediction and management of credit risk has gradually become a hot research topic in recent years, many scholars have also established Logistic model for credit evaluation of small and medium-sized enterprises, but only for the analysis of the real estate industry to establish a model of empirical research is relatively small, and does not fully use non-financial data to establish a model. Therefore, the analysis and research on the credit risk of the real estate industry based on multi-source data fusion can overcome the limitations of the traditional credit risk index evaluation method only based on financial indicators to a certain extent, and can strengthen people's attention to non-financial data in credit evaluation, and establish a credit evaluation model with higher accuracy and stronger feasibility.

2. Literature Review

Foreign economists started their research on credit risk measurement and assessment relatively early. The early credit assessment system mainly assessed corporate credit based on financial indicators. William Berver (1966) assessed the accuracy of credit risk assessment based on financial indicators through the financial indicators of 79 companies with higher probability of default and 79 companies with lower probability of default matched with them in terms of industry and scale during the decade 1954-1964 in the United States. Altman(1968) studied the loan cases of banks in previous years, obtained data from them, conducted research and analysis using mathematical statistics, and assessed the credit risk of borrowers using financial indicators. Edimster (1972) established the enterprise credit evaluation index system, but in this enterprise credit evaluation index system only

contains financial indicators. Haldeman (1977) et al. extended the credit evaluation model under the previous research results, and established a credit risk evaluation model with wider application scope under this premise. Then, Ohlson (1980) used Logistic regression analysis to construct an enterprise credit risk assessment model under the premise of more relaxed assumptions, and used this method to predict the enterprise credit risk. Compared with the previous model, the accuracy of the prediction model is improved.

Many scholars began to find that if only financial data is used to evaluate the credit risk of an enterprise, it cannot meet the expectation well. Therefore, they think that besides financial data, the external macroeconomic conditions of an enterprise will also affect the credit risk of an enterprise. Therefore, gradually some scholars will fully consider the impact of some external environmental factors on the enterprise credit risk assessment. Rose, Andrews and Girous (1982) believe that macroeconomic variables, such as unemployment and the Dow Jones Industrial Average, can affect corporate operations. Lussier (2001) has established a credit risk prediction model for enterprises. Its selected indicators are relatively novel, including non-financial year indicators such as the level of financial management of enterprises, the stage of economic life cycle in which enterprises are located, and the age of operators. The prediction accuracy of the model has been significantly improved. Hu and Ansell(2005) considered four external variables and constructed an enterprise credit evaluation system based on political environment, economic environment, social and cultural environment and technological environment. Campbell et al. (2008) used the dynamic Logistic regression model to study the credit risk of U.S. enterprises, believing that financial data and stock market data can improve the accuracy of the model for predicting enterprise credit risk. Nikolic et al. (2013) based on the financial statement data, established a Logistic model to forecast the credit risk of the enterprise, and used this model to forecast the credit default of the enterprise. Castillo et al. (2018) measured the impact of principal-agent relationship between lenders and borrowers on credit risk under traditional financial ratios and asymmetric information conditions, in order to measure the default risk of SMEs. ZhuHaoliang(2019) based on the real estate development business as a starting point, expanded the current situation of commercial banks' real estate credit, summarized the relevant impact parameters, and verified and explored the credit risk model in the real estate industry. Qing Guan An (2019) combines financial indicators with non-financial indicators, uses principal component analysis and Logistic regression to establish a credit risk identification model, and uses the established model to calculate the probability of future default of small and medium-sized enterprises.

Compared with foreign countries, the theoretical research on credit risk started late. By industry, Gao Fang (2016) based on the Logistic model, carried out empirical research on three different industries respectively, studied and analyzed the fitting degree of credit risk assessment between different industries and Logistic model, and finally found that Logistic model of manufacturing credit risk was the most fitting. The credit risk of real estate industry is one of the hot topics in domestic research in recent years. From the perspective of model data, Li Huiping (2017) uses the theoretical method of system dynamics to establish a risk assessment model for the real estate supply chain, and puts forward suggestions on reducing illegal operations and carefully selecting pledge. Wan Xiangyu (2020) has verified the effectiveness of Logistic model for credit risk assessment of real estate industry, which has certain enlightenment for personnel engaged in this industry.

3. Credit Risk Related Theories and Models

3.1. Basic Theory Of Credit Risk

Credit risk refers to the risk that the counterparty cannot fully perform the contract, which is also called the risk of default. It refers to the risk that the debtor will not perform its obligation to repay the debt, causing losses to the creditors. Credit risk is a common risk in financial risk.

Specifically for domestic listed companies, credit risk mainly includes the financial risk of insolvency of the company and the operating risk of impairment of the company due to lack of management. Listed companies regularly disclose corporate reports in accordance with the

requirements. According to the rules, the stock exchange marks "ST" on companies in abnormal financial position.

3.2. Bredit Evaluation Method

There are two main credit risk analysis methods:

(1) Z-Score model

This model is proposed by Altman(1968), which assesses and predicts the probability of financial failure or bankruptcy of an enterprise through five traditional financial indicators. Generally speaking, the lower the Z value, the higher the probability that the relevant enterprise will have a default event; On the contrary, the higher the Z value, the lower the probability of the relevant enterprise having a default event.

(2) Logistic model

Logistic model uses financial ratio variables and other qualitative variables to predict the probability of corporate default, and determines the risk range based on the risk preferences of financial institutions and investors. Under this condition, the default probability of the company is predicted and analyzed. The following rules can be used to judge the enterprise default risk: if the probability value is higher than 0.5, the enterprise default risk is higher; If the probability value is less than 0.5, it can be determined that the enterprise has a low risk of default.

3.3. Logistic Model

The essence of Logistic model is to find the relationship between independent variables and dependent variables and obtain the probability of events. The probability of default is assumed to be p, p $\in$ 0,1. The function is:

$$P = Pr(y_t = 1|x_t) = \frac{e^{x_t'\beta}}{1+e^{x_t'\beta}} = \frac{1}{1+e^{-x_t'\beta}} \quad (1)$$

Where Pr (…) represents the probability of y=1, and its value range is between 0 and 1; Y is a linear combination of independent variables$(x_1, x_2, \dots, x_n)$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

Among them, is a constant term, is the estimation coefficient of the independent variable, and the function form of the Logistic model is as follows:$\beta_0 \beta_1, \beta_2, \dots, \beta_n$

$$\Pr(y_t = 1|x_t) = \frac{1}{1+\exp[-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)]} \quad (3)$$

In the empirical study, if the value of P tends to 0, it indicates that the enterprise has good credit and its default risk is small; If the value of P tends to 1, it indicates that the enterprise has poor credit and has a high risk of default.

4. Real Estate Industry Credit Risk Empirical Research

4.1. Samples and Data

The relevant data in this paper are from the small and medium-sized enterprise board, GEM, Shanghai A-share and Shenzhen A-share enterprises in Guotai 'an Database, which can completely cover all types of enterprises in the real estate industry. According to the classification of listed companies by CSRC, ST and non-ST companies are determined as low-risk and high-risk companies respectively, and then Logistic model is used to process the data to verify the classification results. In this paper, 123 companies in the real estate industry from 2019 to 2020 in Guotai 'an database are selected as research samples.

4.2. Principal Component Analysis

In order to prevent the correlation between data from interfering with the model construction, principal component analysis is used to process the financial data. Before extracting principal component factors, the data should be standardized.

As shown in Table1, the KMO value of the selected sample data is 0.609, which is greater than 0.5, indicating that the variables selected in the research process of this paper are suitable for principal component analysis. The processing results showed that the 21 indexes passed the requirement of 0.01 significance level and could be factor analyzed.

Table 1 KMO and Bartlett Tests

KMO sampling suitability quantity.		.609
Bartlett sphericity test	Approximate chi-square	1515.387
	freedom	210
	significance	.000

4.3. Principal Component Analysis and Extraction

In order to obtain principal component factors which can reflect more data information from the new variables, the gravel map is drawn and selected from the map. The importance of each factor can be observed as shown in Figure 1. The larger the eigenvalues corresponding to different factors, the more data information they reflect. They should be selected as a new variable to build the model.

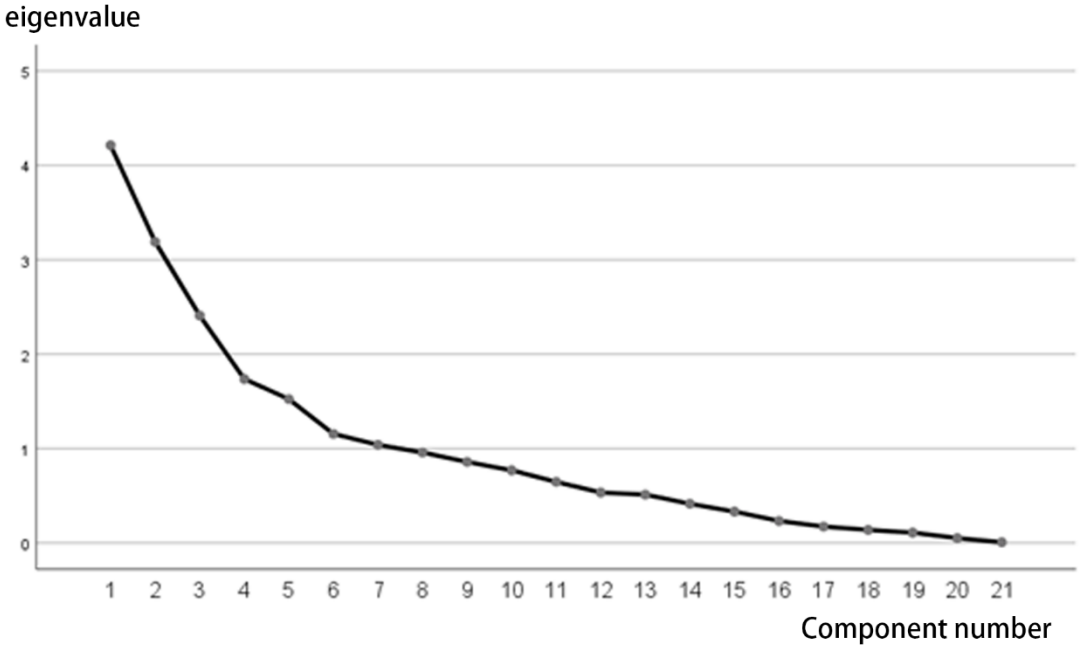

Figure 1 Gravel Chart

According to Figure 1, it can be seen that the top seven factors have the highest eigenvalues, indicating that these seven factors reflect the most data and information, and it is not difficult to find that, starting from the eighth factor, the eigenvalues of the factors become smaller and smaller. And through the verification of the previous theory, if the characteristic value of the factor is less than 1, the factor cannot be selected to construct the model.

4.4. Model Establishment

4.4.1 Logistic Model Based on Financial Data

According to the analysis results above, the linear relationship between 8 principal components and each index is obtained, and the values of ST enterprise and non-ST enterprise are defined as 1

and 0 respectively, and the binary Logistic regression model is established with default probability as dependent variable and principal component factor as independent variable.

Table 2 Variables in Equation

	B	standard error	Wald	freedom	significance	Exp(B)
Step 0	constant	-2.225	.304	53.594	one	.000	.108

As can be seen from Table 2, the first step of the initial assignment method of the system to the model is to assign only the constant term, and the obtained results are B=-2.225, standard error S.E.=0.304, and Sig.=0.000, reaching a significant level.

Table 3 Explanations of significance of regression equation variables

			score	freedom	significance
Step 0	variable	Y1	25.946	one	0.000
		Y2	8.726	one	0.003
		Y3	9.941	one	0.002
		Y4	0.05	one	0.823
		Y5	0.886	one	0.347
		Y6	1.543	one	0.214
		Y7	2.996	one	0.083
	General statistics		50.087	seven	0.000

Table 3 is a preliminary logistic regression model including only constant term model. The significance level of Score test values of the seven principal components is significantly different, and the fourth, fifth, sixth and seventh principal components do not pass the significance level test of 0.1. In order to make the model more fit, Logistic binomial regression is applied to the selected samples again, and "Forward: LR method" is selected for stepwise regression analysis.

Table 4 Hosmer-Lemelshaw Test

step	chi-square	freedom	significance
one	4.952	eight	.763
two	7.659	eight	.467
three	9.464	eight	.305

By the test of Hosmer_Lemeshow, as shown in Table 4, the final chi-square =9.464, which is less than the degree of freedom 8, the critical value of chi-square with significance level of 0.05 is 15.507, and the final Sig.=0.305, which is much greater than 0.05. Therefore, it can be judged that the model is significant on the whole.

Table 5 List of Hosmer-Lemelshaw Tests

		Probability of default = 0		Probability of default = 1		amount
		measure with instruments	hope	measure with instruments	hope	
Step 3	one	12	11.999	0	0.001	12
	two	12	11.981	0	0.019	12
	three	12	11.939	0	0.061	12
	four	12	11.906	0	0.094	12
	five	11	11.840	1	0.160	12
	six	12	11.747	0	0.253	12
	seven	12	11.647	0	0.353	12
	eight	12	11.355	0	0.645	12
	nine	9	10.818	3	1.182	12
	ten	7	5.768	8	9.232	15

It can be found from Table 5 that the "observed value" of the model is very close to the "expected value" and the overall adaptation effect is ideal, which also confirms the experimental results of the previous tests.

Table 6 Explanations of significance of regression equation variables

		B	standard error	Wald	freedom	significance	Exp(B)
Step 3	F1	-2.227	0.585	14.488	one	0.000	0.108
	F2	-2.855	0.767	13.863	one	0.000	0.058
	F3	-1.230	0.583	4.457	one	0.035	0.292
	constant	-4.321	0.848	25.972	one	0.000	0.013

Table 6 illustrates the significance of variables in this binomial logistic regression model. The coefficients of the three principal component independent variables all passed the wald test at a significant level of 0.05, which fully proves that the three variables have certain explanatory power to the model, and the constructed Logistic model has high significance and can help predict the probability of corporate default. From this we get the Logistic model:

$$ln(\frac{\mathrm{P}}{1-\mathrm{P}}) = -4.321 - 2.227\mathrm{F}_1 - 2.855\mathrm{F}_2 - 1.230\mathrm{F}_3$$

Table 7 Test Results Table can be obtained based on the obtained Logistic regression model

Table 7 Test Results

	measure with instruments		predict		
			Low-risk enterprise	High-risk enterprise	Correct percentage
			0	1	
Step 3	Low-risk enterprise	0	109	2	98.2
	High-risk enterprise	1	8	4	33.3
	Overall percentage				91.8

As can be seen from Table 7, there are 123 sample enterprises, including 12 high-risk enterprises and 111 low-risk enterprises. Logistic model will correctly judge low-risk companies with an accuracy of 98.2% and high-risk companies with an accuracy of 66.7%. The Logistic model has a comprehensive prediction accuracy of 95.1% for the overall experimental samples.

4.4.2 Logistic Based On Non-financial Data

Table 8 Hosmer-Lemelshaw Test

step	chi-square	freedom	significance
1	4.506	8	.809

Taking three financial factors and three non-financial indicators as explanatory variables and corporate credit status as explanatory variables, a Logistic regression model is established. The test values of Hosmer and Lemeshow for the regression model are 4.506 and the value of Sig is 0.809, which shows that the model has a good fit. The final regression model is:

$$ln\left(\frac{P}{1-P}\right) = -6.446 - 2.947F_1 - 3.224F_2 - 1.346F_3 + 0.087F_4 - 5.647F_5 + 0.410F_6$$

Table 9 Explanations of significance of regression equation variable**s**

		B	standard error	Wald	freedom	significance	Exp(B)
Step 1	F1	-2.947	1.089	7.320	one	0.007	0.052
	F2	-3.224	1.242	6.738	one	0.009	0.040
	F3	-1.346	0.816	2.719	one	0.019	0.260
	F4 (shareholding ratio of the largest shareholder)	0.087	0.606	0.020	one	0.886	1.091
	F5 (Number of Employees)	-5.647	6.396	0.780	one	0.027	0.004
	F6 (Manager's degree)	0.410	0.638	0.414	one	0.520	1.507
	constant	-6.446	2.713	5.645	one	0.018	0.002

The results can be obtained according to the obtained Logistic regression model.

Table 10 Inspection Results Table

measure with instruments			predict		
			Low-risk enterprise	High-risk enterprise	Correct percentage
			0	1	
Step 3	Low-risk enterprise	0	110	1	99.1
	High-risk enterprise	1	4	8	66.7
	Overall percentage				95.9

4.4.3 Model Comparison

For the comparison of the prediction effects of the models, this paper refers to the evaluation indexes in relevant literature, and selects the Accuracy, Precision, F1-Score, ROC curve and AUC value to compare the accuracy of the two default prediction models established in this paper.

In the default risk forecast in this paper, there are two kinds of forecast results, ST and non-ST, in fact there are two kinds of ST and non-ST, so there are four possible scenarios. Based on these four cases, a confusion matrix of the predicted results can be established. The numerical value of four cases of confusion matrix is determined by the preset risk threshold, which is set to 0.5 in this paper. In this paper, the accuracy, precision and F1-Score are calculated based on the numerical values of four cases of confusion matrix. The predicted results of the two models are shown in Table 11.

Table 11 Classification Forecast of Two Models

measure	sample	Logistic based on financial data	Logistic based on financial and non-financial data
Accuracy	Within sample	0.9186	0.9593
	Out of sample	0.8943	0.9430
Precision	Within sample	0.6667	0.8889
	Out of sample	0.1	0.375
F1-Score	Within sample	0.4444	0.7619
	Out of sample	0.1333	0.4615

According to the results in the table, the accuracy of Logistic model based on financial data alone is lower than Logistic model based on financial data and non-financial data both within and outside the sample.

In terms of accuracy, the inclusion of non-financial data into the Logistic model improves the forecast results, indicating that the model is more effective.

In terms of accuracy, the result value is not high whether it is based on financial data or non-financial data, but the Logistic model based on financial data and non-financial data is slightly higher than the Logistic model based on financial data only.

As for F1-Score, both within and outside the sample, the Logistic model based on financial data and non-financial data is significantly improved as compared with the Logistic model based on financial data only. The accuracy of Logistic model based on financial data and non-financial data

reaches 0.7619, and the accuracy outside the sample reaches 0.4615, which is much better than the 0.444 and 0.1333 of Logistic model based on financial data only.
Next, the ROC curve and AUC value are discussed and analyzed.

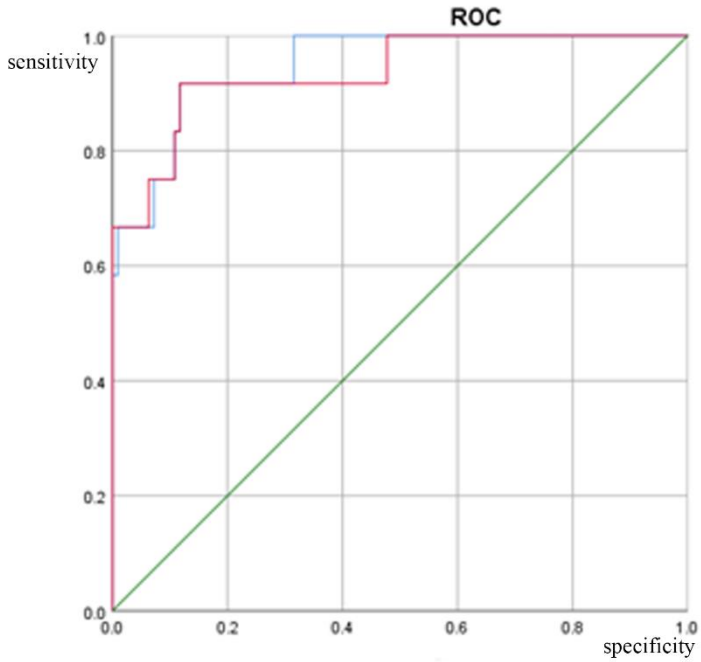

Figure 2 ROC Curve of Two Models

Table 12 AUC Values of Two Models

	zone	Progressive significance
Logistic based on financial data	0.911	0.000
Logistic based on financial and non-financial data	0.948	0.000

In Figure 2, the ROC curve of Logistic model based on financial and non-financial data is more convex upward relative to the ROC curve of Logistic model based on financial data only. According to the results in Table 12, the AUC value of Logistic model based on financial data only is 0.911, and the AUC value of Logistic model based on financial and non-financial data is 0.948. This shows that incorporating non-financial data into the model can improve the prediction accuracy of the model.

4.5. Policy Implications

(1) Strengthen the application of non-financial data in corporate credit evaluation. The evaluation of corporate credit risk should not be limited to modeling and forecasting through financial data, but should combine financial data with non-financial data to build a model that is more accurate and more accurate. Therefore, we should standardize the information disclosure of listed companies, increase the public channels of non-financial data of listed companies, ensure the data are true, objective and accurate, make the information more transparent, and improve the accuracy and stability of the credit risk evaluation model.

(2) Listed companies should strengthen corporate governance and expand the size of the company. Non-financial data in corporate governance information have a crucial impact on corporate credit risk. Therefore, in order to keep the company's credit good, we must try our best to improve the level of corporate governance. At the same time, on the premise of stable development of the company, we can also maintain the company's low credit risk by expanding the size of the company.

5. Summary And Outlook

5.1. Summary of Research

With the further deepening of economic globalization, the ranks of listed companies in China are becoming larger and larger, and the credit risk problem in the real estate industry, which is an important pillar of our economy, has always existed. The default of listed companies has also drawn more and more attention, especially from investors and their relevant stakeholders. This paper selects the financial report data of 123 listed companies in the real estate industry in 2019 and 2020 as the research sample data, and classifies the selected companies as low-risk and high-risk companies according to the non-ST and ST companies. Through the extraction and analysis of financial data and non-financial data, the principal component analysis method is used to extract important indicators, and then the Logistic regression model is used for empirical research. The Logistic model based only on financial data and the Logistic model based on financial and non-financial data are constructed respectively. Finally, the results of the empirical research are compared and analyzed, and a conclusion is drawn.

The study found that non-financial data have a greater impact on corporate credit risk assessment. Among the three non-financial data cited in the paper, the most influential is the number of employees, that is, the more employees, the lower corporate credit risk. The other two non-financial data are related to the corporate governance capability and have certain impact on corporate credit risk. Finally, by comparing the forecast results of Logistic model based on financial data only with Logistic model based on financial and non-financial data, it is found that the forecast results of Logistic model based on financial and non-financial data are more effective and can forecast the credit risk of the company more accurately.

Considering the impact of non-financial data on corporate credit risk, this paper applies the combination of non-financial data and financial data to corporate credit risk assessment, which overcomes the limitations of traditional credit risk assessment methods based only on financial indicators to a certain extent, and can strengthen people's attention to non-financial data in credit assessment, and can make more accurate prediction of corporate credit risk.

5.2. Outlook

When analyzing the problems, the research in this paper has the following deficiencies, which should be improved in the future:

(1) Selected fewer indicators. Although this paper adds three non-financial indicators to the financial indicators selected by the traditional credit evaluation methods, the factors that affect corporate credit risk are far more than these. We should consider the factors that affect corporate credit risk more comprehensively, such as economic environment, social and cultural environment, technological environment and political environment.

(2) The model is not applied to unlisted companies. Due to the limited method of obtaining data, the platform for obtaining and completing the data of unlisted companies cannot be found. The companies studied are all listed companies, which inevitably makes the model limited. In the future, more extensive data should be obtained to improve the model and improve its universality.

(3) The selection of models is less. There are many methods to assess the credit risk of enterprises, but due to time and technology reasons, this paper only applies Logistic model, but it can also use some other methods and models for research, such as Probit model. The Logistic model can be further improved. Compared among different models, the model that best fits the credit risk of Chinese enterprises is selected.

References

[1] Beaver W.H. Financial ratio as predictors of failure[J].Journal of Accounting Research,1966,1(Supplement):71-111

[2] Altman E I . Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. Journal of Finance, 1968, 23(4):589-609.

[3] Edmister R. An empirical test of financial ratio analysis for small business failureprediction[J]J.Journal of Financial and Quantitative Analysis,1972,7(2):147-193

[4] Hamer M. Failure Prediction: Sensitivity of Classification Accuracy to Alternative Statistical Methods and Variable Sets[J].Journal of Accounting and Public Policy,1983:2-11.

[5] Ohlson J A .Financial ratios and the probabilistic prediction of bankruptcy[J]. Journal of Accounting Research,1980,18(1):109-131.

[6] Rose,P.S.,Andrews,W.T.,Giroux,G.A. Predicting business failure: A acroeconomic perspective, Journal of Accounting[J].Auditing&Finance,1982,6(1):20-31

[7] Lussier RN,A Cross-national Prediction Model for Business Success[J].Journal ofSmall Business Management.2001.3:228-239

[8] Hu.Y.C.&Ansell J. Measuring Retail Company Performance by Using Credit ScoringTechniques[M].2005

[9] Campbell J Y , Hilscher J , Szilagyi J . In search of distress risk[J]. Journal of Finance, 2008, 63(6):2899-2939.

[10] Nikolic N , Zarkic-Joksimovic N , Stojanovski D , et al. The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements[J]. Expert Systems with Applications, 2013, 40(15):5932-5944.

[11] Castillo J A , Mora-Valencia A , Perote J . Moral hazard and default risk of SMEs with collateralized loans[J]. Finance Research Letters, 2018, 26:95-99.

[12] Zhu Haoliang. The Research of the Real Estate Credit Risk in Commercial Banks of China[P]. Proceedings of the 2019 International Conference on Economic Management and Cultural Industry (ICEMCI 2019),2019.

[13] Qiguang An,Yuyang Zhao.Construction of Credit Risk Evaluation System for Small-and Medium-Sized Enterprises Based on Principal Component Analysis and Logistic Model[A]. International Science and Culture for Academic Contacts,2019:7.