Journal of Risk Analysis and Crisis Response, 2021, 11(4), 176-188
https://jracr.com/
ISSN Print: 2210-8491
ISSN Online: 2210-8505

Article

Research on Enterprise Credit Risk Prediction Based on Text Information

Haonan Zhang [1,2], Hongmei Zhang [1,2,*] and Mu Zhang [1]

[1] School of Big Data Application and Economics, Guizhou University of Finance and Economics, Guiyang (550025), Guizhou, China
[2] Guizhou Institution for Technology Innovation & Entrepreneurship Investment, Guizhou University of Finance and Economics, Guiyang (550025), Guizhou, China
* Correspondence: zhm1035@qq.com; Tel.: +86-0851-88510575

Received: September 3, 2021; Accepted: December 26, 2021; Published: January 25, 2022

Abstract: This paper uses the text data mining method to separate the intonation in the annual reports of credit risk enterprises and non-credit risk enterprises, quantify it, and study the impact of annual report intonation on the effectiveness of credit risk prediction. In the empirical research, this paper uses the factor analysis method for some traditional financial variables, and uses the extracted components and intonation variables to predict the credit risk through the logistic model. The results show that the tone of enterprises with credit risk is more negative, and the degree of pessimism is significantly positively correlated with the probability of credit risk. By comparing the ROC curves of the prediction results before and after the addition of intonation variables, adding intonation variables to the credit risk prediction based on financial variables can improve the effectiveness of the prediction.

Keywords: Credit Risk; Text Data Mining; Factor Analysis; Logistic Model; Text Intonation

1. Introduction

Credit is the foundation of financial development, and credit risk is also an uneasy factor enough to destroy the whole financial system. Preventing and resolving credit risk is a necessary means to maintain social stability and ensure the healthy development of economy. Nowadays, with the rapid development of Finance and the increasingly frequent financial exchanges among social subjects, it also brings complex interest relations. Once a credit risk occurs in a certain interest link, the associated losses will be immeasurable. Therefore, scholars at home and abroad regard the prevention of credit risk as an important research object.

Credit risk usually refers to the default caused by the reluctance or inability of the borrower, securities issuer, or transaction party to perform the contract [1]. Yang Lian and Shi Baofeng [2] introduced the focal loss modified cross entropy loss function into the credit risk evaluation model to predict the risk of several individual samples. The empirical results show that this prediction method can improve the identification ability of difficult samples. Wang Chongren and Han Dongmei [3] proposed a Bayesian parameter optimization method and XGboost algorithm for personal credit risk assessment of Internet credit industry. The empirical results show that this method is superior to traditional prediction models such as support vector machine. Luo Fangke and Chen Xiaohong [4] brought the Internet Financial personal microfinance data of commercial banks

DOI: https://doi.org/10.54560/jracr.v11i4.311

into the logistic model to screen out the factors that have a significant impact on credit risk. As companies have greater influence than individual borrowers and investors, and the harm caused by credit risk is more destructive, improving the prediction accuracy of corporate credit risk is also a hot issue in the field of risk management. Zhang Tong and Chi Guotai [5] empirically analyzed the data of 2169 Chinese A-share listed companies from the perspective of credit characteristics, and concluded that the model of feature Division has higher discrimination accuracy. Compared with credit characteristics, more scholars' research on credit risk is based on the perspective of the optimal combination of credit risk indicators. Zhou Ying and Su Xiaoting [6] found that different financial indicators have different effects on the prediction of long-term and short-term default status. Li Zhe and Chi Guotai [7] screened 31 indicators with strong ability to distinguish default status from 610 indicators by using the data of listed companies. Li Meng and Wang Jin [8] investigated the impact of enterprise internal control level on its debt default risk through traditional financial indicators. The results show that enterprises with high internal control quality tend to have lower debt default risk.

Previous studies mostly focused on the analysis of financial data. With the progress of computer technology and the rapid development of the Internet, more and more unstructured data are applied to the research of financial problems [9]. Structured data is field variable data. For example, Wu Fei et al. [10] collected the keywords related to "digital transformation" in the enterprise annual report through crawler technology to describe the intensity of enterprise digital transformation. Li Bin et al. [11] identified 29 important risk points in the insurance industry by mining 1682 financial report texts of listed insurance companies in the United States, and analyzed the change trend of important risks in the insurance industry. Liang Kun and He Jun [12] believes that text information effectively alleviates information mismatch and significantly improves the predictability of credit evaluation model. Therefore, text big data can also be applied to the field of credit risk. Cecchini M, et al. [13] extracts the effective information of the management analysis and discussion module in the annual report, and integrates other financial data to improve the prediction default accuracy of the traditional prediction model. Liu Yishuang and Chen Yiyun [14] studied the relationship between text emotion and financial distress through the management tone in the company's annual report. Wang Xiaoyan et al. [15] constructed a priori word frequency of credit risk indicators by mining the text information in journal papers. The empirical results show that the classification effect of credit risk model is significantly improved after using such a priori word frequency. Wang Z, et al. [16] and others believe that in addition to the traditional hard information, soft information can also enter the loan decision-making process, and the effect of credit risk assessment is significantly improved after adding semantic indicators. Zhang Yiwei and Gao Weihe [17] took the borrower's SMS data as the text mining object, analyzed the relationship between the expression of "我" and "我们" and default, and found that the cultural level adjusted the role of these two words in credit risk prediction. Wang Shuxia et al. [18] identified the characteristics of the lender from the text description and used these characteristics to evaluate the credit risk of the loan. The empirical results show that the text data can effectively replace the traditional financial data, and the combination of structured data and unstructured data can improve the performance of the credit risk evaluation system.

It can be seen from the existing literature that the research on the use of text information for credit risk prediction at home and abroad mostly focuses on individual investors, while the research on corporate credit risk mainly focuses on traditional structured data, but there is also a lot of information in many public information such as the company's annual report. Obtaining this kind of text information will help us reduce the impact caused by information asymmetry, so as to improve the effectiveness of credit risk prediction. In view of this, this paper will combine the real default data, select 25 listed companies with credit risk and 53 listed companies without credit risk from 2018 to 2020 as the total sample, peel the tone from the company's annual report, conduct quantitative analysis, and predict the company's probability of credit risk combined with traditional structured data. The research value of this paper is to expand the traditional credit risk identification indicators, prove that it is meaningful to add intonation to the risk identification model, provide new ideas for predicting credit risk, increase China's identification means of credit risk, and enhance the monitoring of systemic financial risk.

2. Research Methods and Index Selection of Enterprise Credit Risk Identification

2.1. Logistic Model and Research Ideas

Logistic regression is a common machine learning method, which is mainly used to classify samples and belongs to "generalized linear regression". This model is often used in credit risk research, such as Zhang Jie and Zhang Yuansheng [19], Bian Yuning et al. [20], Liang Weisen and Wen Simei [21], etc. The reason is that the logistic model has the excellent characteristics that the value of dependent variable is between 0 ~ 1 and does not need to obey normal distribution [22]. The expression of logistic model is:

$$\ln(P/1-P) = \beta_0+\sum\beta_i*X_i \tag{1}$$

In this paper, the enterprise with credit risk is marked as 1. In formula (1), P represents the probability of credit risk, β_0 is a constant term, X_i is a dependent variable affecting the predicted credit risk, β_i is the influence degree of each dependent variable on credit risk. The research idea of this paper is as follows: Firstly, the dimensionality of multiple financial indicators is reduced, and three main components are extracted by factor analysis method. Secondly, the logistic model is used to predict credit risk in two steps. In the first step, only three principal components are input to predict credit risk, and in the second step, three principal components and intonation variables are used as input data to predict credit risk. Finally, the ROC curve is used to compare the credit risk prediction effect of the model before and after adding intonation variables, and the BP neural network model is used to test the robustness of the empirical results.

2.2. Data Selection

Since the financial status and annual report of the enterprise in the year of credit risk will not be known to investors in that year, the annual report and financial data of the year before the occurrence of credit risk are the main basis for investors to predict whether the enterprise has credit risk. When selecting the data of defaulting enterprises, the company's annual report and financial data of the year before the occurrence of credit risk are selected as risk identification indicators. When selecting the data of non-defaulting enterprises, the 2019 annual report and financial data are uniformly selected as risk identification indicators. The annual report data are from the public disclosure of listed companies on Shanghai Stock Exchange and Shenzhen Stock Exchange, and the financial data are from the RESSET financial research database.

2.3. Data Processing

Financial data processing: referring to the selection method of financial indicators in the construction of credit risk identification system by Wang Qianhong and Zhang Min [22] and Liu Xiangdong and Wang Weiqing [23], this paper divides the traditional financial indicators into 5 primary indicators and 12 secondary indicators, as shown in Table 1. Due to the vacancy value in some original financial data, this paper fills it with the average value of this index.

Table 1. Credit risk identification index.

Primary index	Secondary index	Symbol
Debt service level	Quick ratio	X1
	Asset liability ratio	X2
Profit level	Operating profit margin	X3
	Net profit margin on sales	X4
	Return on assets	X5
Operational capability	Turnover rate of fixed assets	X6
	Total asset turnover	X7
	Turnover rate of noncurrent assets	X8
Cash flow indicators	Cash content of operating income	X9
Growth index	Growth rate of net assets	X10
	Growth rate of total assets	X11
	Growth rate of main business income	X12
Unstructured indicators	Annual report intonation	TONE

Quantitative processing of text intonation: This paper uses the "dictionary model" to construct the text intonation of the annual report, and refers to HowNet DICTIONARY [13] and actual financial terms as the emotion dictionary. The dictionary is divided into positive emotion dictionary and negative emotion dictionary. When quantifying the text intonation, first convert the format of the company's annual report downloaded by Shanghai Stock Exchange and Shenzhen Stock Exchange, convert the PDF file format into TXT file format, and then use the Jieba word segmentation package in Python to segment the annual report [24]. Then remove the stop words such as "的" and "了", and make word frequency statistics according to the emotional dictionary. The statistical method is as follows: If there are words in the negative emotion dictionary in the annual report, such as "怀疑", " 难", "疑惑", etc., sum the occurrence times of such words, and use neg to represent the total occurrence times of negative words in an annual report. If the words in the dictionary of positive emotion appear in the annual report, such as "奖励", "引领", "支持", etc., sum the occurrence times of such words, and use POS to represent the total occurrence times of positive words in an annual report. Since negative intonation often has a greater impact on decision makers [14], this paper quantifies the text intonation with formula (2), in which the meanings of NEG and POS have been introduced above. Tone indicates the text intonation, that is, the larger tone, the stronger the negative emotion revealed in the text, otherwise it indicates that the text intonation is more positive. In order to facilitate readers to intuitively understand the positive and negative words in the annual report, this paper generates word clouds from the high-frequency words in the two types of words, as shown in Figure 1 and Figure 2.

$$\mathbf{TONE = NEG/(POS+NEG)} \tag{2}$$

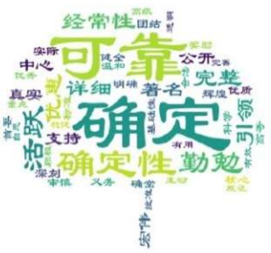

Figure 1. Positive word cloud.

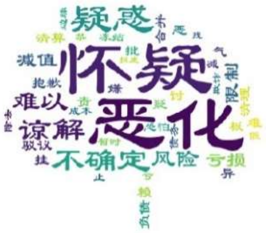

Figure 2. Negative word cloud.

3. Empirical Analysis

3.1. Descriptive Statistics and Inter Group Difference Test

This paper uses SPSS 21 software to carry out descriptive statistics and mean "independent sample t-test" on the data of credit risk group and non-credit risk group, and observe the characteristics of the two groups of data and whether there is significant difference [24]. The enterprise with credit risk is marked as 1 and the enterprise without credit risk is marked as 0. The descriptive statistics of main variables and the results of "independent sample t-test" are shown in Table 2.

It can be seen from table 2 that the *p* value of four variables X6 (turnover rate of fixed assets), X8 (turnover rate of noncurrent assets), X9 (cash content of operating income) and X10 (growth rate of net assets) is greater than 0.05, that is, the difference of these four indicators is not significant and cannot better reflect the difference between different types of samples. The other 9 variables including tone passed the "independent sample t-test", which proved that the remaining 9 variables could significantly reflect the differences between groups. In addition, through the analysis of descriptive statistics, the average, maximum and minimum values of tone of enterprises with credit risk are significantly higher than those without credit risk, which indicates that negative emotions are widespread in the annual report of enterprises one year before credit risk.

DOI: https://doi.org/10.54560/jracr.v11i4.311

Table 2. Descriptive statistics of main variables and t-test results of independent samples.

Explanatory variable	Enterprise defaults	Minimum	Maximum	Mean	Standard deviation	*p* Value
TONE	0	0.02263	0.04906	0.03484	0.00662	0.031**
	1	0.02733	0.06407	0.03873	0.00857	
X1	0	0.27020	4.82480	1.38805	1.17206	0.007***
	1	0.08130	3.30150	0.68852	0.64901	
X2	0	12.68890	88.53180	47.17486	21.25192	0.000***
	1	30.96460	175.83540	74.88301	26.74978	
X3	0	-18.53640	36.24070	8.94938	10.60676	0.001***
	1	-440.15130	52.32540	-35.86637	92.32086	
X4	0	-15.05380	32.30110	7.52735	8.59006	0.001***
	1	-502.99920	40.38250	-43.95637	108.17100	
X5	0	-3.67370	16.37470	5.30242	4.43122	0.000***
	1	-109.20290	12.94880	-8.63132	26.38773	
X6	0	1.16790	102.52400	8.30737	14.71136	0.762
	1	0.34160	94.96940	9.50070	19.00566	
X7	0	0.06230	1.22330	0.60662	0.30374	0.006***
	1	0.05700	1.30740	0.39472	0.32921	
X8	0	0.09130	4.86020	1.66720	1.09202	0.375
	1	0.11960	7.39710	1.37499	1.78710	
X9	0	30.15580	132.85340	99.72378	18.74031	0.110
	1	65.39690	186.61100	108.65078	29.70843	
X10	0	-11.74240	26.17940	5.85995	8.03406	0.093*
	1	-136.35060	257.63410	-10.39964	68.99186	
X11	0	-15.97900	34.99220	6.18474	8.52548	0.009***
	1	-75.05330	46.43540	-3.97918	24.77621	
X12	0	-59.68250	156.73350	9.72063	26.17668	0.022**
	1	-68.08130	63.29080	-6.46469	33.30012	

*** indicates $p<0.01$; ** $p<0.05$; * $p<0.1$.

3.2. Factor Analysis

Based on the theoretical and practical impact analysis, the correlation test is conducted for the variables that pass the "independent sample t-test". It can be obtained from table 3 that tone has no significant correlation with the financial variables, but the correlation between the financial variables is relatively significant, which indicates that there may be some same information between the variables and can be explained to each other. If all variables are input into logistic model to predict credit risk, it may lead to wrong conclusions. Therefore, this paper makes factor analysis on the other 8 variables except.

This paper uses SPSS 21 software to conduct factor analysis on 8 financial variables. Firstly, the data are processed by Z-score standard method to eliminate the influence of sample data dimension [25]. It can be seen from table 4 that the KMO test value is 0.549, which is greater than the standard value of 0.5 and the *p* value is 0. Again, the eight financial variables contain more similar information and are suitable for factor analysis.

Table 3. Correlation coefficient of each variable.

	TONE	X1	X2	X3	X4	X5	X7	X11	X12
TONE	1.000								
X1	-0.070	1.000							
X2	0.066	-0.633***	1.000						
X3	-0.107	0.154*	-0.406***	1.000					
X4	-0.047	0.079	-0.376***	0.971***	1.000				
X5	-0.062	0.115	-0.637***	0.480***	0.537***	1.000			
X7	0.140	0.103	-0.154*	0.173*	0.184*	-0.023	1.000		
X11	-0.070	0.056	-0.385***	0.336***	0.320**	0.677***	-0.098	1.000	
X12	-0.113	0.030	-0.166***	0.412***	0.377***	0.282**	0.166*	0.366***	1.000

*** indicates p<0.01; **p<0.05; *p<0.1.

Table 4. KMO and Bartlett Test.

Kaiser-Meyer-Olkin		0.549
Bartlett's sphericity test	χ^2	436.809
	df.	28
	***p* Value**	0.000

Table 5. Explains the total variance.

Ingredients	Initial eigenvalue			Extract sum of squares load		
	total	variance %	Accumulate%	total	variance %	Accumulate%
1	3.404	42.550	42.550	3.404	42.550	42.550
2	1.366	17.075	59.626	1.366	17.075	59.626
3	1.242	15.527	75.153	1.242	15.527	75.153
4	0.833	10.413	85.566			
5	0.670	8.381	93.947			
6	0.326	4.080	98.027			
7	0.140	1.750	99.778			
8	0.018	0.222	100.000			

Table 6. Composition matrix.

Variable	Ingredients		
	1	2	3
X1	0.340	-0.834	0.149
X2	-0.725	0.599	0.056
X3	0.834	0.267	0.282
X4	0.825	0.316	0.267
X5	0.803	-0.013	-0.394
X7	0.187	-0.072	0.773
X11	0.650	0.110	-0.548
X12	0.533	0.352	0.118

It can be seen from table 5 that three principal components with eigenvalues greater than 1 are extracted from factor analysis, and the three components contain 75.153% of the total variables. The original financial variable data can be reduced by nearly two-thirds through factor analysis, indicating that the result of factor analysis is good. Let the extracted three components be F1, F2 and F3 respectively (See table 6). The scores of each variable in F1, F2 and F3 are shown in table 7. The following expressions are listed according to the scores.

DOI: https://doi.org/10.54560/jracr.v11i4.311

$$F1 = 0.1X1-0.213X2+0.245X3+0.242X4+0.236X5+0.055X7+0.191X11+0.157X12 \quad (3)$$

$$F2 = -0.610X1+0.439X2+0.196X3+0.231X4-0.01X5-0.052X7+0.081X11+0.257X12 \quad (4)$$

$$F3 = 0.12X1+0.045X2+0.227X3+0.215X4-0.317X5+0.622X7-0.441X11+0.095X12 \quad (5)$$

Table 7. Component score coefficient matrix.

Variable	Ingredients		
	1	2	3
X1	0.100	-0.610	0.120
X2	-0.213	0.439	0.045
X3	0.245	0.196	0.227
X4	0.242	0.231	0.215
X5	0.236	-0.010	-0.317
X7	0.055	-0.052	0.622
X11	0.191	0.081	-0.441
X12	0.157	0.257	0.095

3.3. Logistic Regression

When using logistic regression for the first time, only F1, F2 and F3 are used as input variables. The results are shown in Table 8. According to the prediction results of the model, equation (6) can be obtained.

$$Logistic\ (P) = -2.46F1+0.638F2-0.709F3-0.995 \quad (6)$$

Table 8. Logistic regression without TONE variable.

Explanatory variable	Coefficient	Standard error	Wald	df
F1	-2.460***	0.683	12.963	1
F2	0.638	0.533	1.434	1
F3	-0.709	0.508	1.947	1
Constant	-0.995***	0.373	7.123	1

*** indicates p<0.01; **p<0.05; *p<0.1.

Table 9. Logistic regression with TONE variables.

Explanatory variable	Coefficient	Standard error	Wald	df
TONE	0.774**	0.343	5.090	1
F1	-2.628***	0.705	13.909	1
F2	0.693	0.584	1.412	1
F3	-1.015*	0.578	3.080	1
Constant	-1.108***	0.396	7.832	1

*** indicates p<0.01; **p<0.05; *p<0.1

According to the logistic regression results including principal components F1, F2 and F3, only F1 is significant at the level of 1%, and the pre F1 coefficient is -2.460, which is negatively correlated with the probability of default. According to formula (3), the variables with higher scores in F1 are X2 (asset liability ratio), X3 (operating profit margin), X4 (net profit margin on sales) and X5 (return on assets), of which X3, X4 and X5 are variables reflecting profitability, and the coefficient of such variables in F1 is positive. Therefore, it can be inferred that the profitability of a company is the main factor affecting its default. When the profitability of the company is good, credit risk is not easy to occur. The worse the profitability, the higher the probability of credit risk.

When using logistic regression again, input the three components F1, F2 and F3 together with tone variables into the logistic model to predict credit risk. According to table 9, the logistic expression (7) with tone can be obtained, where P represents the probability of occurrence of credit risk.

$$\text{Logistic (P)} = 0.774\text{TONE}-2.628\text{F1}+0.693\text{F2}-1.015\text{F3}-1.108 \tag{7}$$

According to the logistic regression results with tone variable, the pre tone coefficient is 0.774, which is significant at the level of 5%, indicating that tone is significantly positively correlated with the probability of credit risk. According to equation (2), the larger the value of tone, the more negative the tone in the annual report, that is, the more pessimistic the tone in the annual report of the enterprise in the previous year, the more likely the company is to have credit risk, on the contrary, it is less prone to credit risk. It is worth mentioning that after the tone variable is added, the F3 component has changed from having no significant impact on the credit risk prediction in the first regression to being significant at the 10% level. The variables with higher scores in the F3 component are X7 (total asset turnover rate) and X11 (total asset growth rate). Because this is not the focus of this paper, this phenomenon has not been analyzed in detail.

3.4. ROC Curve Comparison

When most scholars use logistic model to predict credit risk, they usually use out of sample resampling method, take the occurrence probability of credit risk of 0.5 as the critical value of risk occurrence, and judge the accuracy of the model to predict credit risk [22, 23, 25], but few articles discuss the scientificity of the critical value. Therefore, this paper uses ROC curve to study the effectiveness of credit risk prediction before and after adding tone to logistic model. The ordinate of ROC curve represents sensitivity, and the higher the index, the higher the diagnostic accuracy; the abscissa represents 1-specificity. The lower the index, the lower the misjudgment rate. Therefore, in general, the closer the point to the upper left corner of the coordinate, the better the diagnostic effect, that is, the larger the area at the lower right side of the ROC curve, the better the credit risk prediction effect.

Table 10. Area under ROC curve

Test result variable		**TONE prediction probability**	**Non-TONE prediction probability**
Area under curve		0.892	0.855
Standard error		0.036	0.049
Significance		0.000	0.000
Asymptotic 95% confidence interval	**Lower limit**	0.821	0.760
	Upper limit	0.963	0.950

DOI: https://doi.org/10.54560/jracr.v11i4.311

Figure 3 shows the comparison of ROC curves of logistic model for predicting the occurrence probability of credit risk before and after the addition of tone variables. The blue curve represents the prediction probability curve of credit risk with tone, and the green curve represents the prediction probability curve of credit risk without tone. It can be clearly seen that the blue ROC curve is closer to the upper left of the coordinate than the green ROC curve. According to table 10, the area under the ROC curve with tone variable prediction results is 0.892, which is greater than the area under the ROC curve without tone prediction results by 0.855. Both results show that adding intonation to predict enterprise credit risk can improve the effectiveness of credit risk identification.

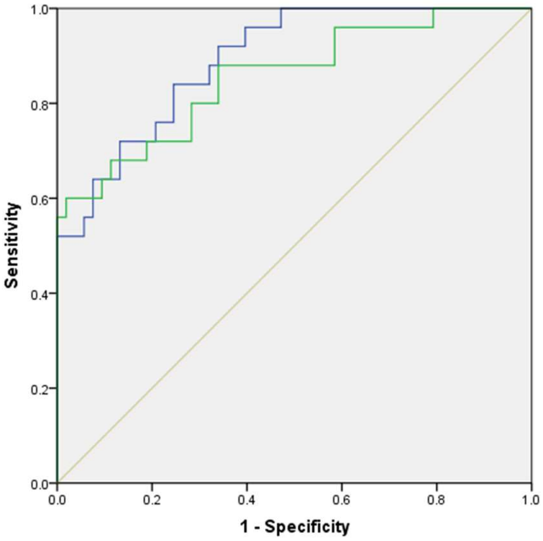

Figure 3. ROC curve.

4. Robustness Test

This paper uses BP neural network model to test the robustness of the empirical results. Firstly, it forecasts the credit risk of traditional financial variables, sets the maximum number of iterations for 10000 times, and trains the total samples with 30% test set and 70% training set. Through multiple training comparisons, the model is optimal when there are 12 neuron nodes, and the results are shown in table 11.

Table 11. Identification results of 12 nodes without tone samples in BP model.

Predicted correct number	48
Prediction errors number	30
Correct rate	61.5%
Error rate	38.5%

Table 12. Identification results of tone samples at 13 nodes in BP model.

Predicted correct number	63
Prediction errors number	15
Correct rate	80.1%
Error rate	19.9%

Then, the traditional financial variables are combined with tone to predict the credit risk. The maximum number of iterations, the total number of samples, and the distribution ratio of training set

DOI: https://doi.org/10.54560/jracr.v11i4.311

and test set remain unchanged. Through comparison, the model is optimal when there are 13 neurons, and the results are shown in Table 12.

From the comparison of the two output results, the number of correct predictions increased significantly after adding intonation variables, and the prediction accuracy increased from 61.5% to 80.1%, an increase of 18.6%, indicating that adding intonation variables can improve the accuracy of model prediction, which is consistent with the empirical results.

5. Conclusions

Taking 25 listed companies with credit risk and 53 listed companies without credit risk from 2018 to 2020 as the research object, this paper uses text data mining method to capture and quantify the intonation, and uses factor analysis method to extract three principal components from the traditional financial data. Finally, we compare the impact of logistic model on the accuracy of credit risk prediction before and after adding intonation variables, and draw the following conclusions. First, when using traditional financial data to predict credit risk, profitability has a great impact on credit risk prediction. The stronger the profitability, the less prone to credit risk. Second, the tone of the annual report of enterprises with credit risk in the previous year is more pessimistic than that of enterprises without credit risk. Investors can observe the tone of the company's annual report to reduce the impact of information asymmetry. Third, according to the empirical results, in the logistic regression, the probability of credit risk is significantly positively correlated with the pessimistic degree of the quantified text tone at the level of 5%, that is, the more negative the tone of the enterprise in the previous year, the greater the probability of credit risk in that year. This result also shows that the tone of the company's annual report contains information related to credit risk, which can solve the problem of information asymmetry between investors and company subjects to a certain extent. Fourth, the ROC curve is used to test the prediction results of the logistic model twice. The results show that compared with the logistic model which only uses the traditional financial data as the input, the effectiveness of the model prediction is improved after adding the text intonation index. It also shows that although enterprises can beautify financial data and increase investor confidence, the negative emotions revealed in the annual report are widespread. By mining the text information of the annual report, we can expand the credit risk identification indicators and improve the effectiveness of credit risk identification.

This paper separates the intonation from the company's annual report, quantifies it, and supplements the traditional credit risk identification system based on structured data. According to the research conclusion of this paper, the most of investors, commercial banks and other financial institutions should strengthen the acquisition of text information when predicting enterprise credit risk, build a risk prediction system from multiple dimensions, improve the efficiency of credit risk identification and reduce the loss caused by information asymmetry.

Funding: This research was funded by the Regional Project of National Natural Science Foundation of China, grant number 71861003 and Guizhou University of Finance and economics is in urgent need of the support of the special subject (2020ZJXK20).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] Chen Yanli, Jiang Qi. Business environment, real earnings management and credit risk identification [J]. Journal of Shanxi University of Finance and Economics, 2021, 43(09): 98-110.

[2] Yang Lian, Shi Baofeng. Credit risk evaluation model and empirical evidence based on Focal Loss modified cross entropy loss function [J/OL]. Chinese management science: 1-12 [2021-10-10]. https://doi.org/10.16381/j.cnki.issn1003-207x.2020.2188.

[3] Wang Chongren, Han Dongmei. Internet credit personal credit evaluation based on hyper-parameter optimization and integrated learning [J]. Statistics and decision-making, 2019, 35(01): 87-91.

[4] Lou Fangke, Chen Xiaohong. Credit Risk Assessment and Application of Individual Microfinance Based on Logistic Regression Model [J]. Financial Theory and Practice, 2017, 38(01): 30-35.

[5] Zhang Tong, Chi Guotai. Default Discrimination Model Based on Optimal Credit Feature Combination - A Case Study of Chinese A-share Listed Companies [J]. System Engineering Theory and Practice, 2020, 40(10): 2546-2562.

[6] Zhou Ying, Su Xiaoting. Enterprise credit risk prediction based on optimal index combination [J]. Journal of System Management, 2021, 30(05): 817-838.

[7] Li Zhe, Chi Guotai. Research on Credit Risk of Listed Companies Based on Maximum Index Discrimination and Optimal Relative Membership [J]. China Management Science, 2021, 29(04): 1-15.

[8] Li Meng, Wang Jin. Internal control quality and corporate debt default risk [J]. International finance research, 2020(08): 77-86.

[9] Shen Yan, Chen Yun, Huang Zhuo. The application of text big data analysis in economics and finance: a literature review [J]. Economics (quarterly), 2019, 18(04): 1153-1186.

[10] Wu Fei, Hu Huizhi, Lin Huiyan, Ren Xiaoyi. Digital Transformation of Enterprises and Capital Market Performance - Empirical Evidence from Stock Liquidity [J]. Managing the World, 2021, 37(07): 130-144+10.

[11] Li Bin, Wang Yinghui, Zhu Xiaoqian, Li Jianping. Identification and evolution analysis of important risk points in insurance industry - Based on text risk information disclosed in financial report [J/OL]. System engineering theory and practice: 1-15 [2021-10-10]. http://kns.cnki.net/kcms/detail/11.2267.n.20210528.0838.002.html.

[12] Liang Kun, He Jun. Analyzing credit risk among Chinese P2P-lending businesses by integrating text-related soft information [J]. Electronic Commerce Research and Applications, 40. https://doi.org/10.1016/j.elerap.2020.100947.

[13] Cecchini M, Aytug H, Koehler G J, et al. Making words work: Using financial text as a predictor of financial events [J]. Decision support systems, 2010, 50(1): 164-175.

[14] Liu Yishuang, Chen Yiyun. Management Tone and Credit Risk Early Warning of Listed Companies - Based on Content Analysis of Annual Report [J]. Financial Economics Research, 2018, 33(04): 46-54.

[15] Wang Xiaoyan, Zhang Zhongyan, Ma Shuangge. Credit Risk Assessment Model Based on Text Prior Information [J]. Chinese Management Science, 2021, 29(05): 34-44.

[16] Wang Z, Jiang C, Zhao H, et al. Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending [J]. Journal of Management Information Systems, 2020, 37(1): 282-308.

[17] Zhang Yiwei, Gao Weihe. Self-construction, Cultural Differences and Credit Risk - Empirical Evidence from Internet Finance [J]. Financial Studies, 2020, 46(01): 34-48.

[18] Wang Shuxia; Qi Yuwei; Fu Bin; Liu Hongzhi. Credit Risk Evaluation Based on Text Analysis [J]. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 2016, 10(1): 1-11.

[19] Zhang Jie, Zhang Yuansheng. Research on the measurement of liquidity risk of P2P lending platform based on Logistic - taking Shangjinfu as an example [J]. Friends of Accounting, 2019(21): 124-127.

[20] Bian Yuning, Lu Likun, Li Yeli, Zeng Qingtao, Sun Yanxiong. Implementation of financial venture capital scoring card model based on logical regression [J]. Computer science, 2020, 47(S2): 116-118.

[21] Liang Weisen, Wen Simei. Study on Loan Default Risk Assessment of Small and Medium-sized Agricultural Enterprises - Based on Data of Agriculture, Forestry, Animal Husbandry and Fishery Enterprises in 'New Third Board' [J]. Rural Economy, 2019(11): 93-100.

[22] Wang Qianhong, Zhang Min. Empirical Study on Credit Default Risk Identification of SMEs in China [J]. Shanghai Economy, 2017(01): 91-100.

[23] Liu Xiangdong, Wang Weiqing. Multi-model comparative study on credit risk identification of commercial banks in China [J]. Economic latitude and longitude, 2015,32(06): 132-137.

[24] Zhang Shuhui, Zhou Meiqiong, Wu Xueqin. Annual report text risk information disclosure and stock price synchronicity [J]. Modern Finance and Economics (Journal of Tianjin University of Finance and Economics), 2021, 41(02): 62-78.

[25] Zhang Jingui, Hou Yu. Empirical analysis on credit risk of SMEs based on Logit model [J]. Friends of Accounting, 2014(30): 40-45.