Research on the Application of Data Mining In the Financial Risk Early Warning of Listing Corporation

Lin Wang [a], Ying Liu [b]

School of management, Wuhan University of Technology, Hubei 430070, China

[a]605823636@qq.com, [b]fuxiao50419@163.com

Keywords: CRISP-DM, Data mining; Logistic regression, Support Vector Machine, Decision tree, Neural network, The financial risk warning.

Abstract. Based on the CRISP-DM process, studied the data mining Logistic regression, support vector machines, decision trees, neural network model and their application in the financial risk early warning of listed Companies. According to four models can predict the response rate and chose Logistic regression as the final model of data mining, and in accordance with the actual situation of Chinese listed companies to build the financial early warning system. By choosing the electronics industry listed companies as samples, based on the financial data of normal listed companies and ST companies, and conducting experiments to evaluate the model, the results show that: in the CRISP-DM process, based on Logistic regression data mining technology to establish financial risk prediction model, predicting the correct rate and response rate more than 85%,the stability of the model is higher, and verify the effectiveness of the data mining technology in the financial risk warning.

1. Introduction

With the rapid development of Internet technology and the advent of the era of big data, more and more decisions will be made with analysis processing on the basis of mass data. Data mining is produced under the background of such application requirements and rapidly developed [1-2]. Applying the powerful advantage of data mining technology to the financial risk early warning, building an effective and applicable financial risk early warning model, predicting the future level of financial risk of listed companies accurately according to the financial information indexes have very important practical significance for market participants[3-4]. This paper based on the data mining standard process CRISP-DM, from the perspective of business understanding, data understanding, data preparation, used Logistic regression, support vector machines, decision tree and neural network data mining technology to establish a financial risk early warning model, and made a preliminary analysis of the financial risk and financial crisis prediction on this basis, compared and selected the best prediction model, so as to enhance the accuracy of the financial risk early warning.

2. The financial early warning model of data mining

2.1 Data understanding and preparation.

The study selected abnormal financial position and the special treatment (ST) of the listed company and part of the normal listed companies of the electronics industry in Shanghai and Shenzhen from 2012 to 2015 as research samples. Used data is from sina finance net, net of information of listed companies in China and Beijing university of CCER economic and financial database. Because companies have different years of financial crisis, to increase the data on the number of pens, so we will discuss all the annual data analysis together. After data optimization, the total of data is 667, including 62 companies which have financial crisis and 605 normal companies.

After the variable selection process, the 19 independent variables were selected in this model, including 18 continuous random variables, a categorical variable, and the dependent variable is a binary classification variable. Delete variables which have high correlation with each other to solve the problem of collinear. After comparing with correlation coefficient of dependent variables, delete eps, sales gross profit margin, operating yield, net pre-tax interest rate, continuous yield, pre-tax profit per share, quick ratio, equity/assets the eight variables, finally selected variable parameters are shown in table 1.

Table 1 Eventually modeling variable parameters

Return on equity (after tax)	Operating margin	After-tax net interest rates	Berry ratio
Current ratio	Total debt/total net worth	Total assets growth rate	Debt ratio
Cash flow ratio	Net worth per share	Cross shareholding	

Because in the study the companies which have financial crisis accounts for only 9.3% of the original material, compared with the 90.7% normal companies which don't have financial situation are six to one, without ratio distribution balance, the characteristics of the financial crisis may not be easy to appear, so in order to improve this situation, super sampling analysis was needed to be carried out on the data before the modeling. The proportion of sampling modeling was set as 1:1, 1:2 and 1:3 as well as the original proportion was 62:605. When sampling ratio is 1:1, the number of companies which have financial crisis is 62, normal companies is 62; When sampling ratio is 1:2, the number of companies which have financial crisis is 62, normal companies is 124; When sampling ratio is 1:3, the number of companies which have financial crisis is 62, normal companies is 186; In the original data, the number of companies which have financial crisis is 62, normal companies is 605.

2.2 Model construction and evaluation.

Building financial risk early warning model of system is a data mining model design based on the data mining tools Clementine12.0 of SPSS Company. The tool faces huge amounts of data in data mining and makes data interaction by creating a series of visual graphic interface, which contains six node area: the source data, record processing nodes and variable nodes, graphics processing node, and output model to create node [5]. The tool also integrates a variety of data mining models such as cluster analysis, decision tree, Logistic regression, support vector machine (SVM), neural network in the visual graphic interface. This study used four kinds of model -Logistic regression, support vector machine (SVM), decision tree and neural network, to predict financial risk of listed companies, combined with the four super sampling proportion, so it needed to build 16 different model, and selected a suitable warning model according to the assessment criteria.

In model analysis, the first thing was drawing random training samples and testing samples according to the proportion of 70% and 30% of the proportion sampling data sets (1:1, 1:2, 1:3 and its original ratio), using Logistic regression, support vector machine (SVM), decision tree and neural network four kinds of modeling methods, in each method of each data set, 50 models was set up. After completion of the modeling, we used test data to calculate basic statistical data of model evaluation index of each model, used the appropriate index to select the most reasonable model. The results of evaluating four kinds of model by an average check rate are shown in figure 1. The figure shows that in each sampling proportion, Logistic regression in the financial risk early warning model is the best, with a 1:1 ratio, the check rate of 86.11%. From the results of 1:1 to 1:3 ratio, of 1:3 proportion sampling check rate is less than 1:1 proportion sampling, but the difference is not big. In comprehensive analysis, in order to use more data samples, decrease the random probability, we decided to establish Logistic regression model under 1:3 super sampling ratio to parameterize early-warning financial risk.

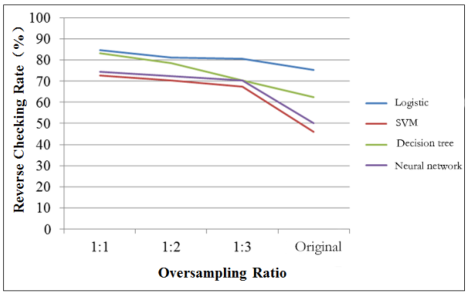

Fig. 1 Model average rate

Comprehensively comparing average check rates of four models above under different sampling exceeding grade, establish Logistic regression model under 1:3 super sampling ratio to parameterize early-warning financial risk. Logistic regression model parameter estimation results are shown in table 2.

Table 2 The Logistic regression model parameter estimation

Variable code	The variable name	B The estimate	S.E.	Wald	df	A significant degree	Exp(B)
Constant		-2.445	0.841	8.451	1	0.004	0.087
F02	Return on equity	-0.038	0.009	19.283	1	0.001	0.962
F12	Current ratio	0.001	0.001	3.966	1	0.045	1.001
F13	Debt ratio	0.052	0.011	24.799	1	0.001	1.054
F15	Net worth per share	-0.091	0.039	5.487	1	0.018	0.913
F17	Cross shares structure	-0.576	0.304	3.575	1	0.059	0.562

To assess model prediction ability, first analyze model of ROC curve, AUC values and Gini coefficient. The area under the ROC curve is defined as the AUC value, indicating the average ability of model discriminating crisis company. If the value is 0.5, indicating to evaluate crisis in the form of random, just like normal probability judgment, there is no difference between abilities, so the AUC value is more close to 1 said discriminant ability is stronger. For Gini coefficient, the calculation formula is 2×（AUC-0.5）, therefore, its representative significance and the AUC value is similar, the closer to 1 said discriminant ability is stronger. The prediction ability is shown in table 3.

Table 3 Model prediction ability

Evaluation of project	The training data	The original material	instructions
AUC value	0.908	0.904	More than 0.7 is better, the more close to 1, the better. The bigger the better.
The Gini coefficient	0.816	0.808	
K - S test value	0.696721	0.684365	

By setting various unfavorable factors, make risk stress tests on the model to evaluate the stability of the model. Risk stress tests, basically can be divided for sensitivity analysis and scenario analysis. Situational analysis mostly need professionals with financial risk management background analyze in view of the current events and the overall financial condition, therefore this study does not discuss it and focus on sensitivity analysis. This study mainly discusses among selected variables which is more sensitive to the variables, namely fix the rest variables below the average parameters while adjust certain variables to the situation of the worst crisis generated by the probability of situation, if there are dramatic changes , it expresses sensitive variables, the results of the analysis as shown in table 4.

It can be concluded from table 4 that when at the best level, the probability of the crisis is 0, when the worst levels the probability of the crisis is 0.99, and when at average, the probability of the crisis was 0.20. In various factors, F02 (return on equity) and F13 (ratio) are sensitive variables, have the maximum impacts on model.

The last is the classification matrix of the original data, as shown in table 5, comparing overall accuracy rates and the check rate of the two, it can be found that the raw data's overall accuracy is 85.87%, the response rate is 84.46%. Because the financial risk was predicted based on sample data, there is always the possibility of miscalculation, so in the model accuracy evaluation, it should cause enough attention to classification error's first category mistake (mistake for ST company not ST company) and the second type of error (mistake not ST for ST), especially, the first category mistake will make creditors or investors make the wrong decision, so as to face huge losses, and the probability of this kind of mistake is higher, therefore it should try to control the first kind of errors. Logistic regression model's probability dividing point is between [0, 1], when the probability of point is not the same, the probabilities of prediction model results making two types of errors are also different. The decrease of the type of error means the increase of the other type of error. Comprehensive formula proposed to determine the optimal threshold by Anderson and Theofanis's control two kinds of wrong theory [6], in this study, default threshold was set to 0.3, under this threshold, it was concluded that the accuracy is 85.87%.

Table 4 Sensitivity analysis results

Project	Coefficient of the model	Optimal levels	The worst level	The average	Single factor changes to worst levels				
					F02	F12	F13	F15	F17
Constant	-2.45	-2.45	-2.45	-2.45	-2.45	-2.45	-2.45	-2.45	-2.45
F02	-0.04	85.69	-198.2	-2.2	-198.06	-2.22	-2.22	-2.22	-2.22
F12	0.01	1948.8	6.01	217.59	217.59	6.00	217.59	217.59	217.59
F13	0.05	1.82	99.76	41.64	41.64	41.64	99.76	41.64	41.64
F15	-0.09	80.52	0.06	15.37	15.37	15.37	15.37	0.06	15.37
F17	-0.58	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PD values		0.00	0.99	0.20	0.99	0.17	0.84	0.50	0.20

Table 5 Classification matrix of the raw data

The actual value	Predictive value	
	Did not produce financial crisis probability (%)	Produce financial crisis probability (%)
Did not produce financial crisis probability (%)	86.08	15.92
Produce financial crisis probability (%)	17.54	82.46
Accuracy (%)	85.87	

2.3 Arrangement and execution.

This study which is under the CRISP - DM process, based on data mining financial early warning model developed by Clementine12.0, set the three parts: model building, model assessment and implementation based on financial risk management system for enterprise users with statements. The interface of this warning model is simple to operate, and can update and adjust the model at any time, in order to ensure the effectiveness of the model.

3. Conclusion

In this paper, developed a financial early warning model analysis system based on the data mining tools Clementine12.0, and applied huge amounts of data mining technology to financial risk analysis, and made test and analysis on our country electronic industry data. The accuracy is more than 85%, the stability is high, and it has a good effect. In the actual evaluation of the listed company financial risk, it often spends considerable time cost in the model, and timeliness will disappear. In this system model, the enterprise users only need to input the relevant financial information and statements, then it can quickly and effectively evaluate the possibility of financial crisis to reduce the cost of the financial risk assessment and judge quickly. Warning model can also monitor the company's financial situation at any time to prevent the occurrence of financial crisis. Today in the European and American countries, most Banks have adopted the technology of data mining to make risk warning, and monitor the financing and investment process of enterprise to prevent malicious commercial fraud, maintain their own interests. With the rapid development of economy in our country, in the context of the Internet +, financial risk early warning analysis which based on data mining technology will also get rapid developments.

References

[1]. Yun Ao. Financial analysis of big data era [J]. Modern business, 2016 (5): 149-150

[2]. Danqing Du. Retail market structure changes of big data era-based on the thinking of electric business enterprise scale expansion [J]. Business economics and management, 2015 (02) : 12 to 17.

[3]. Data mining application in the financial crisis early warning and risk control research [D]. Suzhou: Suzhou University, 2015.

[4]. Jian Li. Listed company financial risk early warning research based on the data mining [D]. Xi 'an: Xi 'an University of Science and Technology, 2009.

[5]. Ja Ma. Research on Data mining technology in the application of the credit risk of listed companies in China [D]. Xi 'an: Xi 'an University of Science and Technology, 2010.

[6]. Michel Crouhy, Dan Galai, Rorbt Mark. A comparative analysis of current credit risk model's[J].Journal of banking & financial,2000(24):59-117.