IJMEHD
INTERNATIONAL JOURNAL ON MANAGEMENT AND EDUCATION HUMAN DEVELOPMENT

J. Management & Education Human Development
ISSN: ISSN: 2775 - 7765
web link: http://www.ijmehd.com

Research on Enterprise Financial Crisis Warning Based on Multiple Eigenvalue Screening Methods

[1]Yuan Yan, [2]Yangfang Zhang, [3]Ling Zhao

[1]Hunan Automotive Engineering Vocational College, Zhuzhou 412001, China;
[2]Business and Trade, Hunan Automotive Engineering Vocational College, Zhuzhou 412001, China;
[3]Business School, Changsha Commerce and Tourism College, Changsha, Hunan 410000, China

Received: 28/02/2024 Accepted: 11/08/2024 Published: 30/09/2024

***Representative e-Mail:** *412668648@qq.com*

ABSTRACT

There are many factors that affect listed companies being ST. Previous research mainly relies on experience, intuition, historical research results and indicators that have seriously deteriorated, and uses a large number of statistical testing methods to screen out the more important indicators (characteristic indicators) in the model. At present, there are many feature index screening methods in machine learning, different screening methods have different effects, each method has its own advantages and disadvantages. This paper takes China's A-share listed companies as the research background, 48 indicators have been preliminarily determined in 6 aspects of capacity, operating capacity, cash capacity, profitability, development capacity and social responsibility indicators. After statistical analysis of the initial indicators, we have obtained the results that can effectively distinguish the sample enterprises of financial crisis and the sample enterprises of financial health 37 indicators, and then use RF, REF, MIC and Lasso to screen the indicator eigenvalues, obtain the eigenvalues and pictures under the four methods, and then obtain 18 eigenvalues according to the application principle of the comprehensive screening method. Finally, the eigenvalues screened by the five indicators are used as the input variables of the financial crisis early warning model, and are brought into the model for empirical analysis. The study found that the characteristic indicators obtained by the comprehensive screening method have the highest accuracy in predicting financial crisis.

Keywords: *Financial Crisis Early Warning; Comprehensive Characteristics; Integrated Analysis Methods; Statistical Testing Methods.*

I. INTRODUCTION

With the continuous development and expansion of China's reform and opening up, Chinese listed companies have developed very rapidly, and their global market share has continued to rise. Due to the global financial crisis and the impact of the global the COVID-19, companies are facing a more complex and volatile social and economic environment, the risk of bankruptcy of the enterprise is aggravated, therefore, more and more stakeholders of the enterprise are concerned about the financial crisis of the enterprise. Under the policy guidance of the China Securities Regulatory Commission, it has announced a special treatment system (ST) for listed companies whose operating conditions have deteriorated, to remind consumers of the trading risks. There are many factors that affect listed

companies being ST. The previous research mainly relies on experience and intuition, historical research results and indicators of serious deterioration, and uses a large number of statistical testing methods to screen out the more important variables in the model to obtain conclusions. To make up for the shortcomings of existing research, this paper takes China's A-share listed companies as the research background, and hopes to establish an efficient and sensitive financial crisis screening index system, select key indicators and substitute them into the financial crisis early warning model, and finally draw a conclusion. The established financial crisis index system has very important theoretical and practical significance for the prevention and control of enterprise financial crisis risk.

II. RESEARCH METHODS

2.1 Introduction to the Theory of Feature Selection Methods

LASSO method is a compressed estimation. It obtains a more refined model by constructing a penalty function, so that it compresses some coefficients, and even directly turns some coefficients with smaller absolute values to zero, so it is especially suitable for the reduction of the number of parameters and the selection of parameters, so it is used to estimate Linear model with sparse parameters; Recursive Feature Elimination (RFE), which is proposed by Guyon et al. based on support vector machines. It requires an updated importance permutation measure at each step of the algorithm. The main idea of recursive feature elimination is to repeatedly build a model (such as SVM or regression model) and then select the best (or worst) features (which can be selected according to coefficients) to extract them, and then repeat this with the remaining features process until all features are traversed. Random Forest

In machine learning, a random forest is a classifier that consists of multiple decision trees, and its output category is determined by the mode of the categories output by the individual trees. Which main idea of calculating feature importance in random forest is to calculate the average value of the contribution of each feature to each number in the random forest. The Maximum information coefficient (MIC for short) can be used to measure the linear or nonlinear correlation strength between two variables, and the MIC value can reflect the importance of features.

2.2 Integrated Analysis Method

In 2017, Professor Zhou Zhihua of Nanjing University proposed the gcForest (multi-Grained Cascade Forest) method to build a deep forest, which is a non-neural network style deep model. This is a new ensemble of decision trees with a cascade structure that enables forests to perform representation learning. Its representation learning capability can be further enhanced by multi-granularity scanning, enabling gcForest to be context or structure aware. Cascading levels can be determined automatically so that model complexity can be determined in a data-dependent manner, rather than being hand-engineered before training; this enables gcForest to work well even on small scale data and enables users to computing resources control training costs. Also, gcForest has much less hyperparameters than deep neural network dnn. The better news is that its performance is fairly robust to hyperparameter settings; experiments from Prof. Zhihua Zhou show that in most cases, excellent performance can be obtained with default settings, even across different data in different domains. The deep forest (gcForest) algorithm is inspired by deep neural networks and ensemble methods, and it is a new ensemble method based on decision tree-based classification. It mainly consists of two parts, Cascade Forest Structure and Multi-grained Scanning.

III. DISCUSSION

3.1 Sample selection and indicator primaries

This article uses whether China's A-share listed companies are ST as a sign of financial crisis. Through the screening of data, 215 companies that were ST in the four years from 2016 to 2019 were selected as financial crisis sample companies, and 1099 normal companies with the same industry and similar total share capital during the same period were selected as matching companies. At the same time, in order to verify the effective effect of the dimensionality reduction method, it is necessary to select as many indicators as possible in the initial selection. A

Co-responding Author: Yangfang Zhang
Hunan Automotive Engineering Vocational College, Zhuzhou, China

total of 41 financial indicators and 7 social responsibility indicators in Table 1 were selected as the primary selection indicator system.

Table 1 Summary of Initial Indicators

Solvency	Management capacity	Profitability	Cash flow capacity	Development ability	Social Responsibility Index
X1Current ratio	X8Accounts Receivable Turnover Rate	X15 Return on assets	X24Net cash content of net profit	X33Capital preservation and appreciation rate	S1Shareholder contribution rate
X2Quick ratio	X9Inventory turnover	X16Net profit margin of total assets	X25Net cash content of operating income	X34Capital accumulation rate	S2Interest payment rate
X3Cash ratio	X10Accounts payable turnover rate	X17Net profit margin of current assets	X26Net cash content of operating profit	X35ROE growth rate	S3Employee contribution rate
X4operating income	X11Cash and cash equivalent turnover rate	X18Net profit margin of fixed assets	X27Total cash recovery rate	X36Basic earnings per share growth rate	S4Supplier contribution rate
X5Assets and liabilities	X12 Liquid assets turnover rate	X19ROE	X28Operating Index	X37Net profit growth rate	S5Consumer contribution rate
X6Equity ratio	X13Capital intensity	X20EBIT	X29Cash fit ratio	X38Operating profit growth rate	S6Tax contribution rate
X7Operating current debt ratio	X14Turnover rate of total assets	X21Return on invested capital	X30Cash reinvestment ratio	X39Total operating income growth rate	S7Social donation rate
		X22Long-term return on capital	X31Cash meet investment ratio	X40Sustainable growth rate	
		X23return on investment	X32Equity free cash flow	X41Growth rate of net assets per share	

3.2 Statistical Testing of Initial Indicators

For the 47 indicators in the full text, the missing value processing is first performed, and it is found that the missing rate of four indicators X35, X36, X37, and X38 exceeds 20%. Since the excess rate is too large, this paper directly removes these four indicators. In order to ensure the accuracy of the construction of the financial crisis early warning model, it is necessary to ensure that the indicators have significant differences between financial crisis companies and financially healthy companies, so it is necessary to test the significance of indicators. Because the specific distribution of each index determined initially is not clear, according to statistical theory, the correlation test was carried out using SPSS 24.0 software. First, the one-sample KS test is used to test whether the indicators conform to the normal distribution, and then the independent sample T test is used to screen the indicators that obey the normal distribution test.

As can be seen from Table 2 of the above nonparametric test results, at the significance level of 0.05, most of the indicators can distinguish ST companies and normal companies, but the quick ratio (X2), equity ratio (X6), cash equivalent turnover ratio (X11), investment rate of return (X23), equity free cash flow (X32), interest payment ratio (S2), and employee contribution ratio (S3) are greater than 0.05 indicating that they are not significant . Therefore,

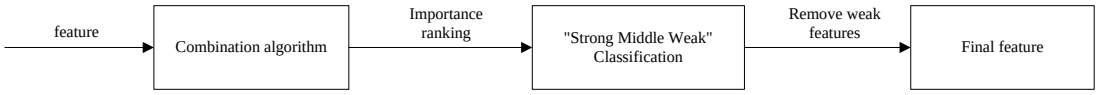
according to the previous statistical analysis theory, these indicators are removed, and the remaining 37 indicators have significant differences and can be brought into the follow-up research.

3.3 Feature selection

This paper constructs a relatively comprehensive index system by selecting financial indicators and social responsibility indicators as the basis for feature selection. The process of selecting indicators is shown in Figure 3-2, the selection steps are as follows:

1. Calculate the feature importance of each method under the four feature methods Lasso, RFE, RF, and MIC separately;
2. According to the tertiles, the feature importance is divided into "strong, medium and weak", and the "strong" importance feature selection is based on: each eigenvalue is calculated, and the eigenvalue greater than the 2/3 quantile is taken; " The selection basis of the "medium" importance feature is: each eigenvalue is calculated, and the eigenvalue between the 1/3 quantile and the 2/3 quantile is selected; the value of the "weak" importance feature is located in the eigenvalue less than the 1/3 quantile, which is the remaining eigenvalues of the first two.
3. The selection principle of combined eigenvalues is: if there are two or more “weak” label values in the classification eigenvalue importance of the four algorithms, delete the feature, otherwise keep it.

Figure 1 Flowchart of selecting indicators



LASSO

Lasso uses a penalty function to compress some of the learned eigenvalues to zero, so as to achieve the function of selecting variables, which plays a role in dimensionality reduction. In this paper, the parameter alpha of the Lasso model is set to 0.01 through five-fold cross-validation, and its feature selection model is established. The output of the model is shown in Figure 3 below. It can be seen from the figure that there are only 8 feature indicators that conform to the indicator screening principle, which are : X13, X29, X9, X10, X24, X18, X26, X28.

RFE

RFE can sort the importance of features to achieve the purpose of feature selection. In this paper, Lasso regression is used as the model basis, and the RFE selection feature model is established through five-fold cross-validation. The output result is the ranking of feature importance. The ranking of feature importance is as follows shown in Figure 4. According to the tertile division method, the first third of the features in the RFE feature importance ranking are defined as "strong" features, and a total of 12 strong features are obtained: X28, X26, X18, X24, X10, X8, X20, X4, X27, X30, X31, X1.

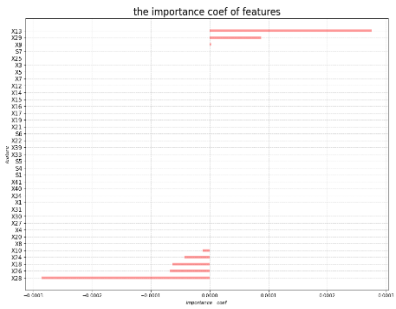


Figure 2 Lasso characteristic coefficient map

Figure 3 RFE feature importance ranking

RF

Random forest has a feature selection function. It can judge how much each feature contributes to each tree in the random forest, then take an average, and finally compare the contribution between features. In this paper, the parameter n=200 of the random forest is determined by five-fold cross-validation, and the RF feature selection model

is established. The output result is the feature importance, and the obtained feature result is shown in Figure 5. According to the previous feature selection criteria, the top third of the feature importance is taken as the 'strong' feature, and 12 strong features are obtained: X20, S6, X27, X39, X41, X4, S7, X33, X40, X34, X30, X8.

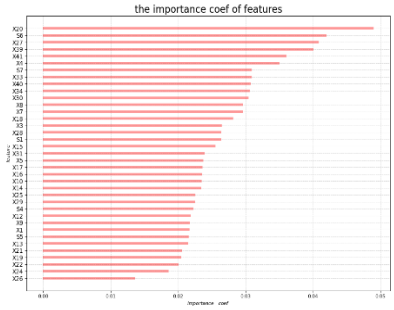


Figure 4 RF Feature Importance

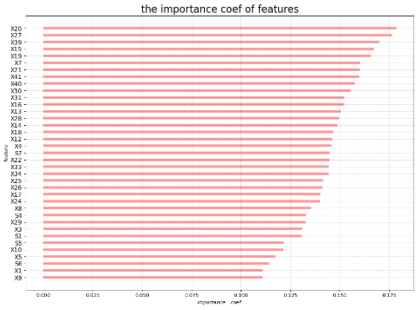


Figure 5 MIC feature importance

MIC

The maximum information coefficient (MIC) can measure the linear or nonlinear association strength between two variables, so feature selection can be performed by calculating the MIC value between variables. The larger the MIC value, the higher the feature importance, and vice versa lower. The model parameters alpha=0.05 and c=10 are set in this paper, and the MIC feature selection model is established, and the feature importance is obtained as shown in Figure 6. According to the tertile division method, the feature whose MIC value is between the two thirds and the maximum value is defined as "strong" feature, then 12 strong features X20, X27, X39, X15, X19, X7, X21, X41, X40, X30, X31, X16.

3.4 Feature Selection Combination Method

Among the 18 feature indicators retained by the combined feature model, the profitability indicators include the net profit margin of current assets (X17), the net profit margin of fixed assets (X18), and the profit before interest and tax (X20); the indicators of operating capacity include: accounts receivable turnover rate (X8), accounts payable turnover rate (X10);indicators of development capability include capital preservation and appreciation rate (X33) capital accumulation rate (X34), total operating income growth rate (X39); debt repayment capability indicators include working capital (X4); cash capacity indicators include net operating profit cash content (X26), total cash recovery rate (X27), operating index (X28), (X29), cash reinvestment ratio (X30), cash satisfy the investment ratio (X31); the social responsibility indicators include shareholder contribution rate (S1) and tax contribution rate (S6). By analyzing the importance of combination features, it is found that most of the indicators are financial indicators, of which the cash capacity indicator is the most reserved, which verifies a proverb in the financial and financial industry "cash is king", cash is like the human body's breathing of air, the flow of blood is so important, most companies go bankrupt, not because of profitability problems, but because of capital chain problems, there is not enough cash flow to deal with the daily expenses of the company, even if good projects are in hand, they have to accept the result of failure. The second is the profitability index. Enterprises can find problems in time through the analysis of profitability index, improve the financial structure of the enterprise, improve the solvency and management ability of the enterprise, and finally improve the profitability of the enterprise and promote the sustainable and stable development of the enterprise. The third is the development ability indicator. The actual data also shows that development ability plays an important role in enterprises. Examining the development ability of enterprises can restrain the short-term behavior of enterprises, which is conducive to improving the modern enterprise system and improving the financial goals of modern enterprises. Compared with other financial indicators, cash capacity, profitability and development capacity have a greater impact on the financial status of Chinese listed companies. Therefore, A-share main board listed companies should pay more attention to and prevent the occurrence of corporate financial crises from the above three aspects. In addition, the innovatively added social responsibility indicators shareholder contribution rate (S1) and tax contribution rate (S6) have obvious importance and have been

retained, which further shows that the non-financial indicators added to this article have a better relationship with the future financial crisis of the company large correlation.

3.5 Prediction results and analysis of different feature selection methods

After the model is established, the model needs to be evaluated. Different types of models have different evaluation criteria. This article is a classification model. The model evaluation methods used include Accuracy, Recall, and measure the effectiveness of the binary model (AUC), Precision. We classify financial crisis samples as (ST) class as positive class, normal company samples as (non-ST class) as negative class, and the 4 cases where the classifier predicts correctly or not on the test set are recorded as: TP—predicts the positive class as the number of positive classes, FN—predicts the positive class as the number of negative classes, FP—predicts the negative class as the number of positive classes, TN—predicts the negative class as the number of negative classes.

$$\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \quad \mathrm{Re}\,call=\frac{TP}{TP+FN}\times 100\% \quad Precision=\frac{TP}{TP+FP}\times 100\%$$

ROC curve and AUC value. The Receiver Operating Characteristic (ROC) curve is a graph with Sensitivity set to the vertical axis and Specificity set to the horizontal axis, which reflects the response of each point on the curve to the same signal stimulus sensitivity. AUC (Area Under Curve) is the integral area under the ROC curve, which obviously will not exceed 1 in value. P-R curve. The P-R curve is the precision vs recall curve, with recall as the abscissa axis and precision as the ordinate axis. When an algorithm classifies a sample, it generally has a confidence level, which means the probability that the sample is a positive sample.

In this section, the four feature selection methods of Lasso, MIC, RFE and RF and the feature selection combination method based on the four methods proposed in this paper are empirically and comparatively analyzed. The above five feature selection methods all use the gcForest integrated model method as the baseline, but the feature index values brought into the model are different. There are 10 feature indicators selected by Lasso; MIC, RFE and RF use the tertile method respectively, and the features with the top 1/3 of feature importance are reserved as the final features; the feature combination is based on the "strong, middle and weak" rule introduced earlier. The features ranked in the top 1/3 of importance are used as the final features, and the experimental results of prediction on the test set are shown in Figure4.

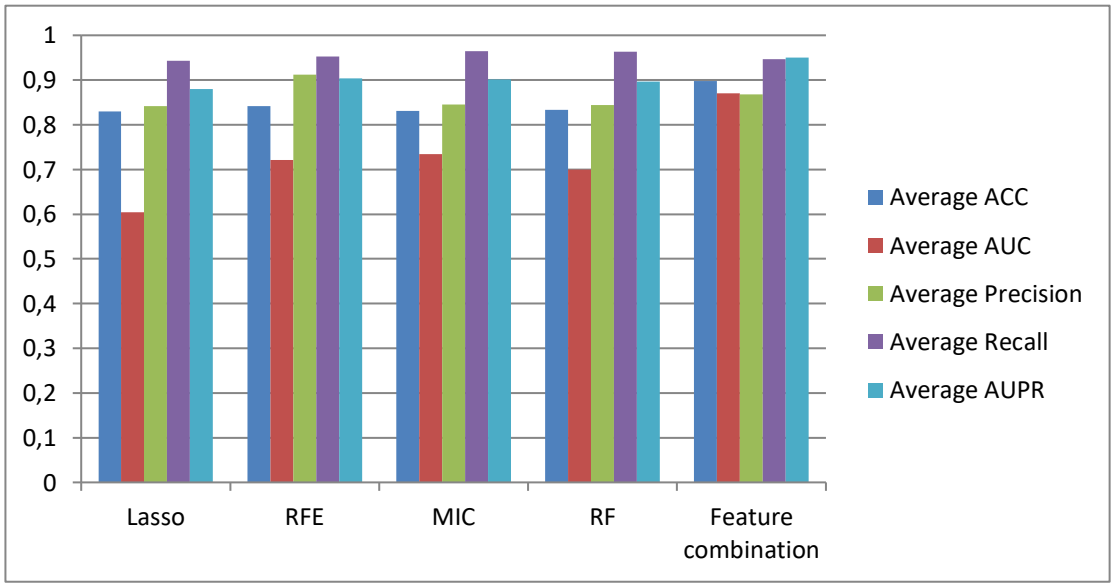


Figure 4 Histogram of screening results of five variables

Through the experimental results, we can find that the feature combination formed by the four basic screening variable methods in this paper has better comprehensive performance than the other four methods. The empirical analysis such as tables and graphs can clearly see that the feature combination method proposed in this paper is effective. Although the precision value of the feature combination method is lower than the RFE and the Recall value

is slightly lower than the MIC, the overall sample accuracy, AUC and AUPR are higher than the other four feature selection methods, which can better reflect the method by reducing variables, It is effective to reuse key variables to reduce computation time and improve the predictive performance of the model.

IV. CONCLUSIONS

This paper takes A-share listed companies as the research object, and preliminarily determines 48 indicators from 6 aspects of solvency, operating capacity, cash capacity, profitability, development capacity and social responsibility indicators, which can effectively distinguish 37 indicators of financial crisis sample companies and financial health sample companies. Then use RF, REF, MIC and Lasso to screen the indicator eigenvalues, obtain the eigenvalues and pictures under the four methods, and according to the principle of the feature combination method, the eigenvalues of 18 feature combinations are obtained, and finally the eigenvalues screened by the five indicators are used as the input variables of the financial early warning model gcForest for empirical analysis. The conclusions are as follows:

In this paper, five variable screening methods of RF, REF, MIC, Lasso and feature combination are brought into the integrated method gcForest. The results show that the model obtained by the method of feature combination variable screening has the highest accuracy, indicating that the feature combination method is suitable for application to the financial crisis warning.

Through the empirical analysis of the importance of feature combination, among the remaining early warning variables, the cash capacity indicator variable retains the most, which verifies the important position of the phrase "cash is king" in the finance and financial industry, and warns listed companies to Pay attention to the company's cash flow management. Through statistical analysis and empirical results, it is shown that the introduced social responsibility indicators have a significant effect on financial crisis early warning.

REFERENCES

Jiao, R., Nguyen, B. H., Xue, B., & Zhang, M. (2023). A survey on evolutionary multiobjective feature selection in classification: approaches, applications, and challenges. IEEE Transactions on Evolutionary Computation.

Kamalov, F., Thabtah, F., & Leung, H. H. (2023). Feature selection in imbalanced data. Annals of Data Science, 10(6), 1527-1541.

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.

Rtayli, N., & Enneya, N. (2020). Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. Journal of Information Security and Applications, 55, 102596.

Samitas, A., Kampouris, E., & Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. International Review of Financial Analysis, 71, 101507.

Wang, P., & Zhang, S. (2015). Correlation analysis method of time-delayed data based on maximum information coefficient. Electronic Measurement Technology(9), 4.

Wang, S., Chen, Y., Cui, Z., Lin, L., & Zong, Y. (2024). Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model. Journal of Theory and Practice of Engineering Science, 4(01), 58-64.

Zhou, Y., Shamsu Uddin, M., Habib, T., Chi, G., & Yuan, K. (2021). Feature selection in credit risk modeling: an international evidence. Economic research-Ekonomska istraživanja, 34(1), 3064-3091.