Journal of Financial Risk Management, 2020, 9, 454-461
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

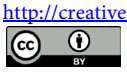
Enterprise Financial Early Warning Based on Lasso Regression Screening Variables

Xi Nie, Guangming Deng*

College of Science, Guilin University of Technology, Guilin, China
Email: *912576024@qq.com

How to cite this paper: Nie, X., & Deng, G. G. (2020). Enterprise Financial Early Warning Based on Lasso Regression Screening Variables. *Journal of Financial Risk Management, 9,* 454-461.
https://doi.org/10.4236/jfrm.2020.94024

Received: October 26, 2020
Accepted: December 5, 2020
Published: December 8, 2020



Abstract

The construction of an enterprise financial warning model is very important for a listed company, and this paper uses the financial data of 2819 listed enterprises as a sample, uses the lasso method for model index screening and uses a variety of classical classification methods and machine learning methods to build the model and analyze its discriminating effect. The results show that the lasso method can effectively reduce the multicollinearity between variables while reducing dimensionality and the classification effect of machine learning method is better than the classical classification method.

Keywords

Financial Early Warning, Lasso Regression, Variable Selection

1. Introduction

Financial risk warning is a process of predicting the likelihood of financial failure of a business and sending warning signals, while it uses a variety of mathematical models to make decisions based on a company's financial statements. The market will give special treatment to listed companies with abnormal financial or other conditions, which are also referred to as ST companies and vice versa as non-ST companies. Bradley Efron et al. (2004) proposed the least angle regression to solve the calculation problem of lasso and promote its popularity in the academic world. Hernandez et al. (2009) proposed the use of lasso to select variables and estimate parameters. Li et al. (2015) Logistic regression was used to construct a corporate financial risk prediction model and analyze the probability of corporate bankruptcy.

The selection of variables and indicators will affect the final model, after reviewing the relevant literature, this paper uses lasso regression correlation algo-

rithm to screen the data variables, combining the classical methods of processing cross-sectional data and machine learning methods to build the financial early warning model and compare the prediction effect of the model through three indicators. Regarding the structure of the article: 1) the paper first introduces the basic theory of the methods and models used; 2) the LASSO method is used to screen variables on real economic data, and then different methods are used to model and compare the data; 3) finally, it is concluded that the Lasso method has good results in dimensionality reduction and that machine learning classification is generally superior to classical classification methods.

2. Lasso-Logistic Model

2.1. Lasso Regression

Assuming that the independent variable data matrix $X=\left\{x_{ij}\right\}$ is an $n\times p$ matrix, ordinary least squares regression seeks those coefficients β that minimize the residual sum of squares. As a method of variable selection, lasso regression requires a penalty term to constrain the size of the coefficient, and ultimately minimize the structural risk and prevent the occurrence of "overfitting". In the case of the penalty term in the constraint condition $\sum_{j=1}^{p}\left|\beta_j\right|\le s$, the coefficient needs to meet the following conditions:

$$\left(\hat{\alpha}^{(ols)},\hat{\beta}^{(ols)}\right)=\arg\min_{(\alpha,\beta)}\sum_{j=1}^{p}\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 \tag{1}$$

Due to the characteristics of absolute value, lasso regression will filter out some coefficients. Mallows C_p is one of the criteria used to evaluate lasso regression. If $p\left(k>p\right)$ is selected from the respective variables of k to participate in the regression, then the C_p statistic is defined as

$$C_p=\frac{SSE_P}{S^2}-n+2p;\ SSE_p=\sum_{i=1}^{n}\left(Y_i-Y_{pi}\right)^2 \tag{2}$$

Based on this, we choose the model with the smallest C_p.

2.2. Logistic Regression

This paper assumes that the dependent variable has two possibilities: the firm is an ST firm or a non-ST firm, which are 1 and 0 respectively. The linear model $Y_i=\beta_0+\beta_1X_1$ does not meet its assumptions in this case, but Y_i is a Bernoulli distribution, so its mean has a special meaning in the model:

$$P=\left(Y_i=1\right)=\pi_i,\ P=\left(Y_i=0\right)=1-\pi_i \tag{3}$$

From this, the Y can be derived:

$$E\left(Y_i\right)=1\times\pi_i+0\times\left(1-\pi_i\right)=\pi_i \tag{4}$$

The π_i in the above formula represents the probability value, which is in line with the basic linear regression, so here you can mostly use logistic regression to fit the model. According to the principle, the following formula is obtained:

$$P_i = f\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in}\right) \tag{5}$$

Y_i can be expressed in another way:

$$P\left(Y_i\right) = \pi_i y_i \left(1-\pi_i\right) 1 - y_i \tag{6}$$

The logarithm of the maximum likelihood function is:

$$LnL = \sum_{i=1}^{n} y_i \ln \pi_i + \left(1-y_i\right) \ln\left(1-\pi_i\right) \tag{7}$$

$$\pi_i = \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in}\right) \Lambda + \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in}\right)$$

Substitute the upper formula to the following equation:

$$LnL = \sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in}\right) - \ln\left[1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in}\right)\right] \tag{8}$$

3. Empirical Analysis

3.1. Sample Selection

All the data in this article are from the CSMAR database, CSMAR database is a research-oriented accurate database in the field of economy and finance, which is based on the professional standards of CRSP, COMPUSTAT, TAQ, THOMSON and other authoritative databases, and is the largest financial and economic database with the most accurate and comprehensive information in China. The data selected the financial data of all 194 ST enterprises (hereinafter referred to as ST enterprises) and 3570 unlabeled ST enterprises (hereinafter referred to as non-ST enterprises) as of September 30, 2019. After processing the missing and abnormal values of the data, the final sample data were 33 labeled ST enterprises and 2786 unlabeled ST enterprises.

3.2. Indicator Description

On the basis of the previous research results, the data variables of solvency, profitability, management ability, development ability and cash flow are selected from five aspects:

- Debt solvency: Reflects the liquidity and debt level of the company's funds, which is conducive to evaluating the company's financial status and financial risks;
- Profitability: profitability is the main goal of enterprise management also reflects the comprehensive ability of the enterprise, the evaluation of the profitability of the enterprise to a certain extent can reflect the financial operation of the enterprise;
- Management ability: reflects the enterprise to the asset utilization and the management situation, to a certain extent can evaluate the enterprise to maintain and increase the value;
- Development ability: reflects the future of the enterprise's gold management

is an important index to predict the development potential of an enterprise;

- Cash flow analysis: dynamically reflects the flow of cash and cash equivalents in a certain period of time. Based on the above considerations, this paper selects 16 indexes around solvency, profitability, operating ability, development ability and cash flow, and draws them into Table 1, 17 indexes as dependent variables and initial independent variables for the financial early warning model of listed companies.

3.3. Introduction of Evaluation Indicators

Most of the samples selected in the papers on the enterprise financial warning model are equal, that is, the number of experimental groups and control groups is the same, so most of them use prediction errors to measure the quality of the model when commenting on the prediction effect of the model. That is, the product of misjudgment and total. However, when the number of different types of variables varies greatly, this evaluation method is not applicable. By consulting the relevant literature, this paper introduces three indexes that can be used to comment on the two categories of variables: accuracy rate, recall rate and F1. Rate.

Table 1. Initial financial warning model indicators.

	Symbol	Indicators	Definition
Dependent variables	x_1	Are ST enterprises	
Solvency capacity	x_2	Current ratio	Current assets/current liabilities
	x_3	Quick ratio	(Current assets − inventories)/current liabilities
	x_4	Property rights ratio	Total liabilities/total owners' equity
Profitability	x_5	Gross assets net profit margin	Net profit/total asset balance
	x_6	Return on net assets	Net profit/shareholder equity balance
	x_7	Net profit/total profit ratio	Net profit/total profit
	x_8	Sales Cost Rate	Sales expenses/operating income
Operational capacity	x_9	Turnover of accounts receivable	Closing balance of operating income/accounts receivable
	x_{10}	Inventory turnover	Operating costs/end of inventory balance
	x_{11}	Total assets turnover	Operating income/total assets closing balance
Capacity development	x_{12}	Growth rate of total assets	(Total assets end of current period − total assets beginning of current period)/(total assets beginning of current period)
	x_{13}	Net profit growth rate	(Net profit current quarter amount − net profit last quarter amount)/(net profit last quarter amount)
	x_{14}	Net asset growth per share	(net assets per share end of current period − net assets per share beginning of current period)/ net assets per share beginning of current period
Cash flows	x_{15}	Cash content of operating income	Cash/operating income received for sale of goods, provision of services
	x_{16}	Net operating income cash	(Net cash flow from operating activities)/(Gross operating income)
	x_{17}	Total cash recovery	(Net cash flow from operating activities)/(Total assets) closing balance

Assuming that the model has four results in prediction, the four results are:
TP: forecast ST enterprises as ST enterprises
FP: forecast non-ST enterprises as ST enterprises
FN: forecast ST enterprises as non ST enterprises
TN: forecast non-ST enterprises as non-ST enterprises
Accordingly, the precision rate P is defined as:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall rates R defined as:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F_1 is the harmonic average of accuracy and recall, defined as:

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

3.4. Model Building

This paper uses the lars package in R software for lasso regression to screen the financial warning model index. Table 2 shows the partial values of C_p statistics in different cases (here only the results of steps 8 to 15). And the minimum value is step 11 ($C_p = 9.5639$), and the variable selection effect is optimal. Table 3 shows the final selection of the software output lasso regression variables are x_3, x_4, x_5, x_8, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}.

Figure 1 gives the increase and decrease of coefficients under the asynchronous number, which can be used to visually determine the selection process of the financial indicator. The left side is the intercept, and the right side is holding all the variables. Figure 1 shows that as the estimated regression coefficients of the variables gradually increase, the coefficients of the different variables show different degrees of dispersion, with the variables showing the largest changes.

Table 4 shows the results of multi-collinearity determination by characteristic root, from which we can see that the number of conditions $k > 100$ before the

Table 2. The change of C_p value of financial data in lasso regression.

step	8	9	10	11	12	13	14	15
RSS	30.98	30.94	30.93	30.80	30.78	30.78	30.77	30.77
Cp value	19.96	18.61	20.12	9.56	10.23	11.67	13.01	15.03

Table 3. Variable selection results.

Variable	x_3	x_4	x_5	x_8	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
Coefficient	0.001	0.011	−0.026	−0.002	0.009	−0.021	0.007	0.003	−0.003	0.005

screened variables in Lasso Regression indicates that there is strong multi-collinearity between variables; the number of conditions $k < 10$ after the screened variables in Lasso Regression indicates that the degree of multi-collinearity between variables is small.

Table 5 shows that the classification methods, in order of F1 value from largest to smallest, are: random forest and adaboost have the highest F1 value of 1; logistic regression, mixed linear discriminant, linear discriminant, and flexible linear discriminant have F1 values of 0.301, 0.232, 0.19, and 0.19, respectively;

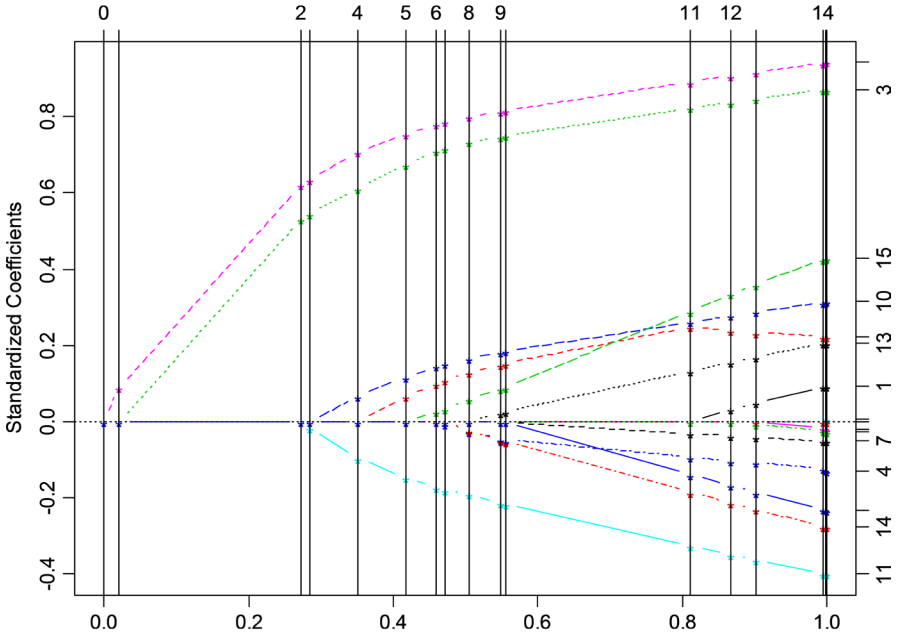


Figure 1. Path map of regression coefficient solution.

Table 4. The result of eigenvalue determination.

	before filter	after filter
Condition number	196.02	3.88

Table 5. Model prediction effect comparison.

	Method	Sensitivity	Recall rate	F1 value	Average F1
Classical Linear Discriminant	Logistic Regression	40%	24.2%	30.1%	22.9%
	Linear Discriminant Analysis	20%	18.2%	19%	
	Mixed Linear Discrimination	22.2%	24.2%	23.2%	
	Flexible Linear Discrimination	20%	18.2%	19%	
Machine Learning	SVM	100%	9.1%	16.7%	57%
	Bagging Classification	100%	6.1%	11.4%	
	Random Forest	100%	100%	100%	
	Adaboost Classification	100%	100%	100%	

the least effective classification is SVM and Bagging classification for 0.167 and 0.114. In summary, it seems that the classification of machine learning methods is generally better than classical methods, and the accuracy of machine learning methods is generally higher than classical methods, but the recall of SVM and Bagging classification is not as high as that of classical methods in F1 value; from F1 value it seems that the best classification among classical classification methods is logistic regression, and its four classification The F1 values of the methods are not as high as those of the machine learning methods overall, but the differences in classification performance between the methods are small.

4. Conclusion

Through the analysis of the financial data of 2819 listed companies as of September 2019, the lasso method is introduced to screen the data index, and the model is established by various classical classification methods and machine learning methods. Finally, the prediction effect of each method is compared by using precision rate, recall rate and F1 value, and the following two conclusions are drawn:

1) The collinearity between variables decreases obviously after the model is screened by lasso method, which indicates that lasso method can effectively reduce the multicollinearity between variables.

2) Taking into account that the collected data is unbalanced (non-ST enterprises account for most of the data), the classification effect of machine learning method is better than that of classical classification method. However, the model of SVM and bagging classification is not as good as the classical classification method.

This paper innovatively introduces the LASSO method in a variety of classical classification and machine learning methods to achieve a better prediction effect with a more streamlined model, which can not only be applied to the classification problem but also extended to the regression problem, and provide readers with a reference when choosing a classification method.

Acknowledgements

This paper is financially supported by National Natural Science Foundation of China (NSFC) under Grant number 71963008.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics, 32,* 407-451. https://doi.org/10.1214/009053604000000067

Hernandez, D. J., Han, M., Humphreys, E. B., Mangold, L. A., Partin, A. W. et al. (2009).

Predicting the Outcome of Prostate Biopsy: Comparison of a Novel Logistic Regression-Based Model, the Prostate Cancer Risk Calculator, and Prostate-Specific Antigen Level Alone. *BJU International, 103,* 609-614. https://doi.org/10.1111/j.1464-410X.2008.08127.x

Li, H. K., Wang, Y., Zhao, P. S. et al. (2015). Cutting Tool Operational Reliability Prediction Based on Acoustic Emission and Logistic Regression Model. *Journal of Intelligent Manufacturing, 26,* 923-931. https://doi.org/10.1007/s10845-014-0941-4