Research on Financial Distress Early Warning of Listed Companies Based on GBDT Model and SHAP

Emma Li[1,a], He YANG[2,b], Jiapei Li[3,c,*], Yifang Cai[2,d]

[1]Henan Experimental High School, Zhengzhou, China

[2]School of Math. And Stats., Zhengzhou University, Zhengzhou, China

[3]Henan Key Laboratory of Financial Engineering, Zhengzhou University, Zhengzhou, China

[a]2607679347@qq.com, [b]2981979556@qq.com, [c]282958703@qq.com, [d]1187659753@qq.com

*corresponding author

Keywords: Financial Distress, Early-warning, GBDT, SHAP

Abstract: Early warning of financial distress is of great significance to the healthy development of enterprises and stakeholders' decision-making. Nowadays, the data-driven machine learning method has replaced the statistical methods to become the mainstream method of financial early warning. However, the complex model brings higher prediction accuracy, and brings the model unexplained embarrassment as well. In this paper, the ensemble learning algorithm GBDT(Gradient Boosting Decision Tree) is introduced to establish a financial distress early warning model, and the interpretability of the model is studied based on the SHAP(SHAPley Additive exPlanations) framework. Then the empirical study on more than 2000 listed manufacturing companies in China shows that the GBDT model has better generalization ability than the model based on Logistic regression, and SHAP gives a clear explanation about how financial distress contributed by feature variables. This study is valuable not only in theory but also in application for financial distress risk early warning.

1. Introduction

With the development of big data and artificial intelligence, machine learning has become one of the main tools for scientific research. The prediction method of financial distress of listed companies has been replaced by data-driven machine learning method from traditional statistical method. Machine learning algorithms, such as Logistic Regression[1][3] and Support Vector Machine[2], have been studied in the financial distress prediction of listed companies, but these machine learning methods are weak learning algorithms. Kearns[4] proposes an ensemble learning approach that "elevates" multiple simple weak learning algorithms to strong learning algorithms. GBDT algorithm, proposed by Friedman[5], is an integrated decision tree algorithm based on gradient lifting. It has been widely used because of its generalization ability and advantages in feature selection. However, while the complex machine learning model brings high prediction accuracy, it loses the interpretability, which brings great obstacles to the application. Fortunately, Lundberg et al.[6] introduced the explanatory principle of SHAP framework for complex models, and concluded that SHAP framework can identify important features and verify them, which has stability, consistency and rationality, and is more in line with people's understanding and judgment conclusions.

Therefore, this paper introduces the GBDT algorithm to establish the financial distress warning model of listed companies. Then the empirical study on more than 2000 manufacturing listed companies from Shanghai Stock Exchange and Shenzhen Stock Exchange is carried out, and the results show that GBDT model has better generalization ability than the traditional Logistic Regression model. And the GBDT model is explained visibly and quantitatively in the framework of SHAP.

  DOI: 10.25236/ecemis.2021.071

2. GBDT Algorithm and SHAP

2.1. Brief Introduction of GBDT

GBDT algorithm is an ensemble learning algorithm composed of multiple decision trees and based on gradient boosting. Given a data set $D=\left\{\left(x_i,y_i\right)\mid x_i\in\mathbb{R}^m, y_i=0\ or\ 1, i=1,2,\ldots,n\right\}$ containing n samples and m features, x_i represents the i-th sample feature and y_i represents the observation category probability value of x_i. Assume that the model requires K iterations in total, that is, K decision trees are built, t is the number of current iteration, $t=1,2,\ldots,K$. The loss function of GBDT classification represents the deviation between the predicted value $\hat{y}_i^{\,t}$ of the model and the true value y_i, which is generally taken as:

$$l\left(y_i,\hat{y}_i^{\,t}\right)=-\left(y_i\log p_i+\left(1-y_i\right)\log\left(1-p_i\right)\right) \tag{1}$$

Where $p_i=\dfrac{1}{1+e^{-\hat{y}_i^{\,t}}}$ represents the predicted probability of x_i being a positive sample. Denote the t-th decision tree generated by the t-th iteration as $f_t\left(x\right)$, then the total loss function of the t-th iteration is:

$$L^{(t)}=L\left(y,f_t\left(x\right)\right)=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{\,t}\right) \tag{2}$$

Where, we need to initialize $\hat{y}_i^{(0)}=0$

Add the first tree to the model: $\hat{y}_i^{(1)}=f_1\left(x_i\right)=\hat{y}_i^{(0)}+f_1\left(x_i\right)$

Add a second tree to the model: $\hat{y}_i^{(2)}=f_1\left(x_i\right)+f_2\left(x_i\right)=\hat{y}_i^{(1)}+f_2\left(x_i\right)$

And so on, we add the t-th tree to the model: $\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k\left(x_i\right)=\hat{y}_i^{(t-1)}+f_t\left(x_i\right)$

It can be seen that $\hat{y}_i^{(t)}$ represents the prediction result of sample x_i from the first tree to the t-tree model. In the gradient boosting algorithm, a weak classifier (i.e. a decision tree) will be generated in each iteration. Assuming that the decision tree obtained in the previous iteration is $f_{t-1}\left(x\right)$, the goal of this iteration is to find a decision tree $f_t\left(x\right)$ to minimize the total loss function $L^{(t)}$ of this round. Gradient boosting refers to that in order to quickly reduce the residual, each training is conducted l in the negative gradient direction with the fastest residual reduction.

2.2. Interpretation Framework Based on SHAP

SHAP is a method framework based on additive feature interpretation. It was first proposed by Shapley in cooperative game theory and used to study the value of each player in cooperative game. Later, it was applied to explain the value of features in complex models. For the output of a model, each feature becomes a participant in the final result of the model. Each sample in the model generates a corresponding SHAP value, which is the sum of the contributions of each feature in the sample to the prediction.

Assume that x_{ij} represents the j-th feature of the i-th sample, $SHAP_i$ represents the predicted contribution value of the i-th sample by the model, and $SHAP_{base}$ represents the baseline of the model, namely the mean value of the target variable of all samples in the model, then the SHAP value satisfies the following equation:

$$SHAP_i=SHAP_{base}+shap\left(x_{i1}\right)+shap\left(x_{i2}\right)+\cdots+shap\left(x_{im}\right) \tag{3}$$

where, $shap\left(x_{ij}\right)$ represents the SHAP value of x_{ij}, and m represents the dimension of the features. If $shap\left(x_{ij}\right)>0$ it indicates that this feature has a positive effect on the predicted value.

If $shap\left(x_{ij}\right) < 0$ it indicates that the feature has a negative effect on the predicted value. The advantage of the SHAP value is that it reflects not only the influence of the features, but also the positive and negative effects of the influence. And SHAP will be mapped to the prediction probability score of the model on the sample by Sigmoid function:

$$\hat{p}_i = \frac{1}{1+e^{-SHAP_i}} \qquad (4)$$

when $\hat{p}_i \geq 0.5$, the sample classification is 1, that is, the positive sample (ST sample); Otherwise the sample classification is 0 and it's a negative sample.

3. Financial Distress Model Based on GBDT

The data in this paper are from CSMAR(China Stock Market Accounting Research) database. The stock of the company marked as ST (including ST, *ST, SST, S*ST and S) is taken as the judgment standard of the financial distress status. If the company is marked as ST, it is considered as suffering from financial distress. Otherwise, the company is considered financially sound. Since a company in 2 years in a row under financial losses is judged as ST, we take the (T-2)-th year data as input data, where T is the year the company began to fall into financial loss. For consistency, we choose the manufacturing industry as study target, from year 2018 to 2020, 2116 companies in the manufacturing industry are non-ST, and the remaining 125 are ST companies.

3.1. Feature System

66 financial features and 14 non-financial features are selected to construct the financial distress early warning model, among which the financial features include seven first-level features, such as operation ability, relative value features, profit ability, development capacity, debt paying ability, a share index and risk level. The non-financial index includes internal governance, equity structure, audit and innovation ability. Specific second-level features are shown in Table 1.

Table 1 Feature system

Index	Code&Name	Index	Code&Name	Index	Code&Name
Financial index					
Operation ability	X_0 ❊: Basic earnings per share; X_1 ❊: Working capital ratio; X_2: Working capital to net assets ratio; X_3: Profit ratio of main business; X_4 ❊: Current asset turnover; X_5: Non-current asset turnover; X_6: Fixed asset turnover; X_7 ❊: Total asset turnover; X_8: Working capital turnover; X_9: Accounts payable turnover; X_{10}: Inventory days; X_{11}: Accounts receivable turnover days; X_{12}: Non-recurring gains and losses; X_{13} ❊: Capital intensity.	Profit ability	X_{23}: Operating margin; X_{24}: Net operating rate; X_{25} ❊: Operating cost ratio; X_{26}: Total operating cost ratio; X_{27} ❊: Return on assets; X_{28} ❊: Total assets profit rate; X_{29}: Return on equity; X_{30} ❊: Cost margin; X_{31}: Expense ratio during sales; X_{32}: Cash fitness ratio; X_{33}: Cash flow adequacy ratio; X_{34} ❊: Total cash recovery; X_{35}: Operating index.	Debt paying ability	X_{46} ❊: Working capital; X_{47}: Current ratio; X_{48} ❋: Quick ratio; X_{49} : Cash ratio; X_{50}: Equity ratio; X_{51} ❋: Equity multipler; X_{52} ❋: Interest earned ratio; X_{53}: Cash flow interest / Matured debt; X_{54}: Long-term debt to capital ratio; X_{55}: Long-term debt / working capital.
Relative value index	X_{14}: Market value; X_{15} ❋: Price earning ratio; X_{16} ❋: Price-to-sales ratio; X_{17} ❋: Price cash flow ratio; X_{18} ❋ : Price-to-book ratio; X_{19} ❋ : Dividend change ratio per share; X_{20} ❋: Cash dividend earned ratio; X_{21}: Earnings retention rate; X_{22} ❋: Tobin Q value.	Development capacity	X_{36} ❊: Growth rate of total assets; X_{37}: Growth rate of return on equity; X_{38}: Growth rate of operating profit; X_{39} ❋: Hedging and proliferating ratios; X_{40}: Basic earnings per share growth rate; X_{41}: Net flow's growth rate per share; X_{42} ❋: Net asset growth per share; X_{43}: Net flow's growth rate from operating activities; X_{44}: Cash flow's growth rate from investment activities; X_{45}: Cash flow's growth rate from financing activities.	A share index	X_{56} ❊: Earnings per share; X_{57} ❋: Net cash flow per share from operating activities; X_{58}: Net cash flow per share from investment activities; X_{59}: Net cash flow from financing activities per share; X_{60}: Gross operating income per share; X_{61}: Operating profit per share; X_{62} ❋: Net assets per share.
Risk level	X_{63} ❊: Financial leverage; X_{64}: Operating leverage; X_{65} ❊: Comprehensive lever.				
Non-financial index					

Internal Governance	X_{66}: Internal control effective or not;X_{67}*: Internal control defect or not;X_{68}: Corrective actions taken or not; X_{69}*: Any major lawsuits or not; X_{70}*: Violate regulations or not.	Equity structure	X_{71} ❊: Z index; X_{72}*: Herfindahl_3 index; X_{73}: Directors' number; X_{74}: Independent directors' number; X_{75}*: Board shares' number.	Audit	X_{76}: Audit opinion type.
Innovation ability	X_{77} ❊ R&D personnel proportion; X_{78}: R&D investment amount; X_{79}: R&D investment to revenue ratio.				

3.2. Feature Engineering

The data processing and feature engineering and modeling in this paper were done in Python3.7.6. First, 3 features with a deletion rate of more than 50% were directly deleted, and the remaining 77 features entered the next stage of screening. Then, for the continuous features, the optimal Chi-Merge chi-square bin separation method is applied. For discrete characteristics, we treat each value of the features as a bin. Then the IV value was used to select features with high predictive ability. When the IV value of the feature was less than 0.08, the predictive ability of the feature was weak and thus was dropped. Then, the remaining 38 features to be selected and enter the next stage. Next, if the correlation coefficient between two features is greater than 0.7, the features with smaller IV value are deleted. At this stage, a total of 2 features were deleted and 36 features were screened out (see Table 1 for features with *). Finally, 36 features were selected to enter the model, where 29 financial features and 7 non-financial features are included.

3.3. Model Construction

This paper adopt SMOTE oversampling method to adjust the ratio of positive and negative samples to 4:1. The data set contains 2616 sample data and 36 columns of features. GBDT function is used for modeling. When training the model, grid search is used to find the optimal parameters for the model by using Pipeline function. Then, the trained model is used to classify and predict the financial status of the test samples, and the final classification results are obtained. The four evaluation measures of GBDT and Logistic models are shown in Table 2:

Table 2 Comparison of model results

	Accuracy	AUC	KS	Kappa
GBDT	0.9044	0.9320	0.7507	0.6738
LR	0.8891	0.9026	0.6602	0.6275

Accuracy represents the ability of the classification model to classify samples correctly, and its value range is between 0 and 1. The higher Accuracy, the higher overall Accuracy of the model classification results. AUC value is commonly used to measure the classification effect of the model, and its value range is [0,1]. The larger the value is, the better the current classification algorithm can classify. KS value is usually applied to the model's ability to distinguish between positive and negative samples, that is, to evaluate the model's ability to distinguish between companies that will be ST in the future and those that will not be ST. Kappa coefficient is an index used to measure the classification accuracy, and its value range is between -1 and 1, but usually the Kappa value falls in the interval [0,1]. The higher the value is, the higher the classification accuracy of the model is. It can be seen from Table 2 that GBDT model is superior to LR in four evaluation measures.

4. Model Interpretation

4.1. Significance and Influence Mode of Features

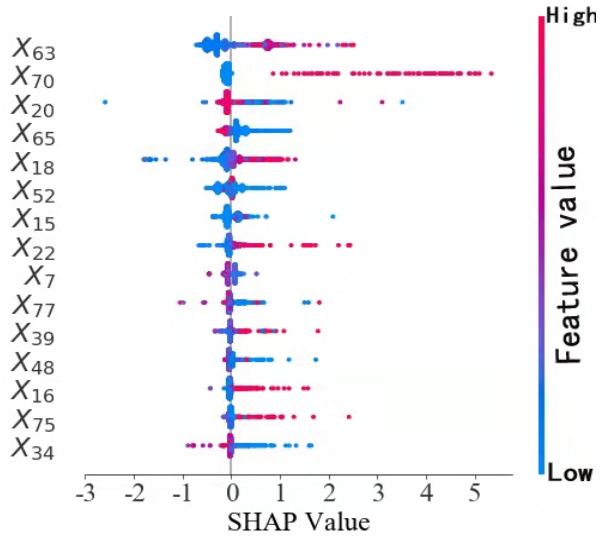

Figure 1 Contribution of each feature to model output

In Figure 1, the left side of y-axis shows the name of each feature, from the top to the bottom, their contribution decreases. And the horizontal axis is the SHAP value. The color on the right side of the graph goes from blue to purple to red, indicating that the feature values are getting larger and larger. It can be seen from the figure that X_{63} makes the largest contribution to the model, and the larger the feature value is, the greater the probability of financial distress occurring, that is, the greater the possibility of being judged as ST. The trend of X_{70} is similar to that of X_{63}. With the increase of X_{70}, the probability of the sample being judged as ST increases, and X_{18} and X_{22} have similar properties. On the contrary, with the increase of X_{65}, the probability of the sample being judged as ST decreases, so does X_{34}.

4.2. Calculation of SHAP Value of a Single Sample

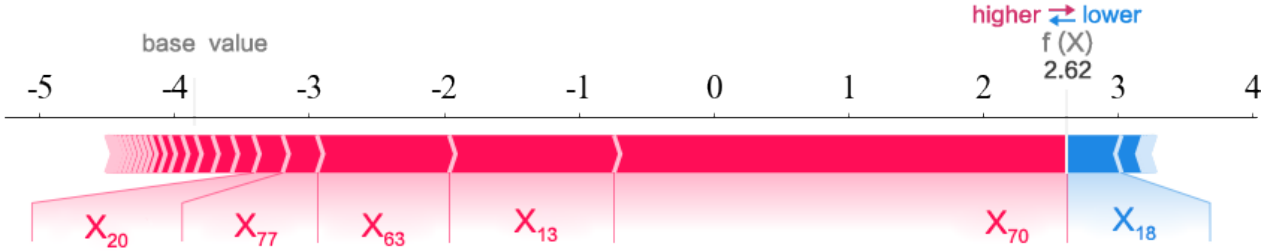

Figure 2 SHAP value of a positive sample

As can be seen in Figure 2, this is the SHAP value of each feature in a positive sample (ST) . The blue feature represents that this feature will weaken the probability of the sample being classified as a positive sample, while the red feature represents that this feature will increase the probability of the sample being classified as a positive sample, and the width of each segment interval represents the influence degree of this feature. For this sample, X_{70} is the most contributing feature, which will increase the probability that the sample is predicted to be a positive sample. The second is X_{13} and the third is X_{63}, and so on, which finally result in a positive sample. It can be seen from Eq.(3) that the sample has a SHAP value:

$$SHAP = SHAP_{base} + shap(X_{70}) + shap(X_{13}) + \cdots + shap(X_{18}) = 2.62 .$$

According to Eq.(4), the probability that the sample is a positive sample is $\hat{p} = \frac{1}{1+e^{-2.62}} = 0.9321$. The sample is judged as ST company because $\hat{p} \geq 0.5$.

5. Conclusion

This paper established a GBDT financial distress early warning model, and studied how to explain the model using SHAP framework. The model introduced 80 features related to financial distress, including both traditional financial features and non-financial features of enterprises. The empirical study was conducted on more than 2000 manufacturing companies listed in Shanghai Stock Exchange and Shenzhen Stock Exchange. The results show that GBDT model has more advantages than traditional Logistic model in terms of AUC, KS and other generalization ability. The importance of features was ranked by SHAP, and the ways how they led to financial distress were given visibly and quantitatively. The research of this paper has important theoretical significance and practical value for enterprises to predict the risk of financial distress.

Acknowledgments

This research is supported by Special Fund for Key Disciplines of Zhengzhou University with funding No. 129/32410380.

References

[1] Han Y., Sun Q., Yu Z. Research on Financial Early Warning of Listed Companies Based on Lasso-logistic Model. Advances in Economics, Business and Management Research, vol.56, pp.262-266, 2018.

[2] Y.F. Li, Q. Zhang. Research on Financial Distress Early-warning of Listed Companies Based on GA-SVM. International Journal of Education and Management Engineering, vol.1, no.2, pp.1-7, 2011.

[3] Lu Xu, Qingzhu Qi & Peiding Sun. Early-Warning Model of Financial Crisis: An Empirical Study Based on Listed Companies of Information Technology Industry in China. Emerging Markets Finance and Trade, vol.56, no.7, pp.1601-1614, 2020.

[4] Kearns M., Valiant L. G. Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing. New York, ACM Press, 1989, pp.433-444.

[5] Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, vol.29, no.5, pp.1189-1232, 2001.

[6] Lundberg S. M., Lee S. I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, vol.30, pp.4765-4774, 2017.