Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33745

Machine Learning-based Financial Risk Early Warning Model for Chinese Listed Companies and an Empirical Study of Influencing Factors

Ran Gu

Department of Mathematics and Applied Mathematics, Shandong University, Weihai, China
202300700045@mail.sdu.edu.cn

Abstract. This paper takes Chinese A-share listed companies as the research background and builds a financial risk early warning model using the XGBoost algorithm in the field of ensemble learning. In view of the extremely imbalanced ratio of ST samples to normal samples in financial data, this study models under the original unbalanced distribution to reflect the early warning behavior in the real market environment. Meanwhile,the SHAP effect attribution framework is introduced to enhance the interpretability of "black box" models, achieving a deep connection between prediction accuracy and financial logic. The empirical results show that this model maintains extremely high prediction accuracy and exhibits significant accuracy characteristics, capable of effectively identifying potential financial risks. The effect attribution analysis further confirms that profitability is the core driving factor for identifying financial distress. This research provides a highly reliable and transparent automated early warning solution for financial regulation and investment decisions, and has important practical reference value for optimizing the risk governance of the capital market.

Keywords: Financial Risk Early Warning, XGBoost, SHAP, Imbalanced Data, Chinese Listed Companies

1. Introduction

Against the backdrop of explosive growth in data volume, the application of data science methods to construct precise classification models from vast financial information has become a core issue in financial risk control. The Special Treatment(ST) system in China's A-share market provides a natural binary classification scenario for financial distress prediction. Building an efficient early warning model not only enhances the level of automated risk identification but also represents a typical practice of data science empowering the financial industry. In this process, handling highly imbalanced datasets (i.e., a very small number of default samples and a large number of normal samples) is a key challenge to improving the practical performance of early warning models [1].

Research methods have shifted from traditional linear models such as Logistic regression to modern machine learning algorithms. In recent years, gradient boosting algorithms represented by Extreme Gradient Boosting(XGBoost) have become powerful tools for processing structured

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.33745

financial data due to their excellent feature extraction capabilities and robustness to missing values. XGBoost, through efficient parallel computing and regularization techniques, can significantly enhance the generalization performance of financial early warning models in complex scenarios [2].

Given this, this paper constructs an XGBoost risk early warning framework based on the CSMAR database. This study does not rely on manual sample balancing techniques and aims to observe the true performance of the algorithm in the original imbalanced distribution. At the same time, the Shapley Additive Explanations(SHAP) attribute framework is introduced to address the issue of interpretability. As a cutting-edge interpretative tool, this framework can effectively break through the trust bottleneck caused by the "black box" nature of machine learning by converting complex decision logic into intuitive feature contribution degrees [3].

This paper mainly makes the following marginal contributions: Firstly, from a data perspective, it explores the classification characteristics of machine learning models under real imbalanced sample distributions, filling the gap in current research that overly focuses on algorithm accuracy while ignoring the influence of the original data structure. Secondly, from a methodological perspective, it combines ensemble learning algorithms with an interpretability framework to achieve a complete data science processing flow from "prediction results" to "logical attribution". Thirdly, it explores the intrinsic influence mechanism of core financial features on the model's determination results from a unique perspective. The method of constructing probability prediction models based on multiple financial indicators lays an important empirical foundation for subsequent risk early warning research [4].

2. Method

2.1. Data sources and description of variables

This study selects Chinese A-share listed companies as the research sample, with financial data and ST labels obtained from the CSMAR database.

2.1.1. Target variable

The label variable mentioned in this article refers to whether the enterprise has undergone special treatment (ST). In the dataset, it is set as a dummy variable named "ST_Label". If a sample enterprise received ST or *ST treatment during the observation year, its value is 1, indicating that the enterprise is in financial distress; if the enterprise's operating condition is normal, its value is 0. This setting transforms the financial warning issue into a binary classification task in data science.

2.1.2. Feature variables

To construct a robust predictive framework, seven pivotal financial metrics were distilled from the dataset, covering four dimensions of corporate health:

Profitability Metrics: This includes ROE and ROA. These two show how well a company earns profit from its assets and equity. They are the primary tools for checking a firm's health.

Solvency Indicators: This paper used the debt-to-asset ratio and the current ratio. The first one checks long-term risk, while the second one shows if a firm can pay its short-term debts.

Operational Efficiency: These are asset turnover and inventory turnover. They show how well management uses its resources. A low rate often means the capital chain is tight.

Growth Potential: This paper used revenue growth to measure this. It shows if a company is still competitive in the market. A steady drop here is often a sign of coming trouble.

DOI: 10.54254/2753-8818/2026.33745

2.2. Preprocessing

This paper cleaned the raw data in two steps to make the signals clearer for the XGBoost model. The main goal was to improve data quality while keeping all important risk information.

2.2.1. Extreme value handling strategy

Financial ratios often show extreme values when the bottom number of a fraction is very small. To handle this, this paper cuts the top 1% and the bottom 1% of all continuous variables. This step is necessary because outliers can confuse the model and distort the SHAP results. By removing these extremes, this paper makes the data more stable for the training process.

2.2.2. Missing value handling strategy

Standard models usually force users to fill in missing data or delete incomplete rows. Table 1 provides the statistical description of variables. This paper avoided these methods because XGBoost can handle missing values on its own. This choice is based on the idea that a gap in a financial report is often a sign of risk. By keeping the missing values, this paper allows the model to learn from these gaps and find hidden patterns in struggling companies. This approach keeps the data authentic and reflects the real world of financial distress.

Table 1. Statistical description of variables

Variable	Sample	Average	Standard	Min	Max
ROE	5570	-0.0051	0.2457	-1.6552	0.2801
ROA	5594	0.0152	0.0725	-0.2951	0.1704
DebtRatio	5594	0.4214	0.2186	0.0570	0.9432
CurrentRatio	5497	2.7022	2.6133	0.3106	15.3087
AssetTurnover	5593	0.5390	0.3738	0.0168	2.2908
RevenueGrowth RevenueGrowth	5452	0.2861	0.8141	-0.7211	5.3850
InvTurnover	5433	15.0500	57.3335	0.2084	496.7048

2.2.3. Sample partitioning and processing

Before model training, the cleaned dataset was randomly divided into a training set and a test set in a 8:2 ratio. Since the proportion of ST samples in the actual market is extremely low, no manual resampling operation was conducted in this study. The purpose was to observe the actual prediction performance of the model under the original, imbalanced data distribution, thereby enhancing the external validity of the empirical results.

2.3. Algorithm selection and model construction

2.3.1. XGBoost model

Due to the non-linear and high-dimensional nature of financial data, this study selected the XGBoost algorithm, which performs well in ensemble learning. XGBoost is an optimized version of the Gradient Boosting Decision Tree (GBDT). By integrating second-order derivative information and regularization terms, it significantly improves the computational accuracy and generalization ability

DOI: 10.54254/2753-8818/2026.33745

of the model, enabling it to effectively handle the complex relationships among financial indicators. In numerous large-scale comparative studies, the algorithm framework has been proven to be robust and computationally efficient when dealing with structured financial data [5].

2.3.2. Model interpretability method (SHAP)

To overcome the limitations of "black box" machine learning models, this study introduced the SHAP framework to conduct attribution analysis of the prediction results. The core concept of SHAP originates from the Shapley value in game theory, and it assigns a contribution score to each feature. Through SHAP analysis, this paper not only obtained the global ranking of feature importance but also analyzed at the individual level how specific financial indicators (such as return on equity or debt ratio) drive the model to classify a company as ST class or normal class, thereby closely linking the prediction results to financial logic. This method provides the necessary transparency for complex credit risk early warning models and represents a key technical means for achieving responsible governance in the fintech field [6].

3. Results

3.1. Evaluation of XGBoost model prediction performance

After completing data preprocessing and feature engineering, the dataset was fed into the XGBoost model for training and testing. To comprehensively evaluate the model's classification performance under extremely imbalanced data, this study not only examined overall accuracy but also focused on precision, recall, and F1-score as core evaluation metrics. Table 2 presents the detailed classification report for the model on the test set.

Table 2. XGBoost model classification report

	precision	recall	f1-score	support
0	0.9491	0.9991	0.9735	1,120
1	0.8571	0.0909	0.1644	66
accuracy	-	-	0.9486	1,186
macro avg	0.9031	0.5450	0.5689	1,186
weighted avg	0.9440	0.9486	0.9284	1,186

3.1.1. Evaluation of overall predictive performance

From a macro perspective, the overall accuracy rate of the XGBoost model reached 94.86%, indicating that the model's prediction coverage across the entire dataset was extremely extensive. This outstanding classification performance has been repeatedly verified in previous studies on credit risk and bankruptcy prediction; especially when dealing with financial indicators with complex interaction relationships, the gradient boosting decision tree algorithm often outperforms traditional statistical models [7]. For normal enterprises (labeled as 0), the recall rate of this model is as high as 99.91%, meaning that in actual enterprises operating normally, the model rarely makes incorrect classifications. This effectively avoids "false alarms" caused by excessive warnings.

DOI: 10.54254/2753-8818/2026.33745

3.1.2. "Defensive" prediction strategy under imbalanced data

A more in-depth analysis of the detailed indicators reveals that this model has significant practical application value in terms of classification features:

For normal enterprises (label 0): The recall rate of this model is as high as 99.91%, which means that in truly operating normally, the model rarely makes incorrect classifications. This effectively avoids erroneous positive results caused by excessive warnings.

For enterprises in financial distress (label 1): The model's predictions exhibit a clear "defensive" characteristic. Its accuracy rate reaches 85.71%, meaning that once the model issues an ST warning signal, the probability that the enterprise is indeed in an ST state is close to 86%, indicating an extremely high credibility of the warning signal. At the same time, its recall rate is only 9.09%. Due to the strict use of real and imbalanced data (without manual oversampling), this model naturally forms a cautious strategy of "prefer to be safe than to take risks": although it cannot capture all potential risky enterprises, any warnings it issues have high reliability. In the actual process of financial risk control and investment decision-making, this highly accurate mechanism can effectively avoid risks and significantly reduce unnecessary portfolio rebalancing costs.

3.2. Feature attribution analysis based on the SHAP framework

To address the "black-box" nature of ensemble learning algorithms, this paper introduces the SHAP framework to provide both global and local visual explanations of XGBoost's decision-making logic.

3.2.1. Identification of core risk drivers

Figure 1 displays the global importance ranking of model features; the length of each bar represents the amount of information contributed by that feature in distinguishing between "normal companies" and "ST companies."

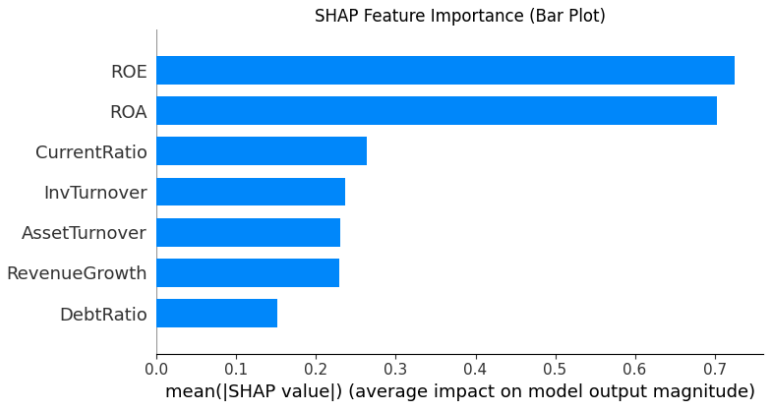


Figure 1. Feature importance ranking (SHAP bar plot) (picture credit: original)

The results show that the return on equity (ROE) and the return on assets (ROA) play an absolutely dominant role. This indicates that profitability is the most important indicator for measuring a company's financial condition. It is particularly important to note that the feature importance ranking output by the XGBoost model represents the "information gain contribution" of each indicator in identifying financial distress. This suggests that profitability is the core "indicator" of financial distress, but it should not be simply equated with a pure causal driving factor.

DOI: 10.54254/2753-8818/2026.33745

3.2.2. Marginal impact mechanism of financial features

In addition to overall importance, this paper further illustrates the specific impact directions (Impact Direction) of each feature variable on the model's output results through Figure 2.

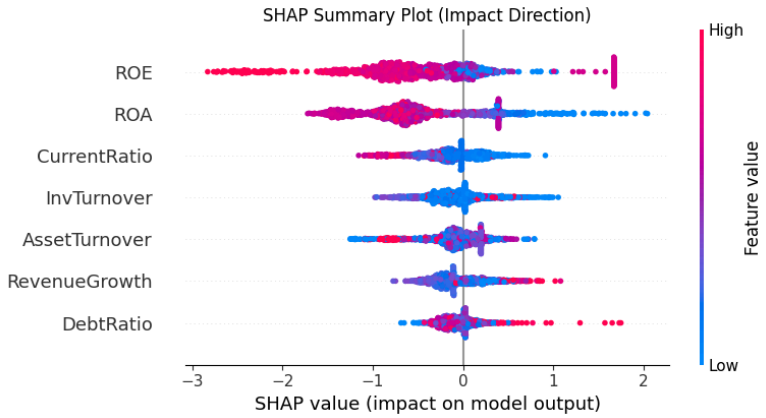


Figure 2. Feature impact direction and distribution plot (SHAP summary plot) (picture credit: original)

Each data point in the Figure 2 represents a sample. The color temperature (warm or cold) indicates the magnitude of the original feature value (red represents a higher value, while blue represents a lower value), and the x-axis represents the size of the SHAP value.

When the red data points mainly cluster on the right side of the y-axis, it indicates that an increase in this indicator value will raise the probability that the model classifies the company as ST;

When the red data points mainly cluster on the left side of the vertical axis, it indicates that an increase in this indicator value has a protective effect on the company (reducing the probability of ST).

Take ROE as an example, the figure shows that the red points mainly cluster on the left side, which means that an increase in the equity return rate significantly reduces the risk of the company becoming ST. On the contrary, if the company's profitability (ROE) is low, it will significantly increase the risk of financial warnings.

4. Discussion

In the tests, the XGBoost model achieved 85.71% precision but only 9.09% recall. This result shows how the model handles rare ST cases in a real, imbalanced dataset. This paper chose not to use resampling tools like SMOTE because this paper wanted to avoid synthetic noise. As Krawczyk explained, models dealing with extreme imbalance often use very strict criteria to prevent wrong guesses [8]. Large-scale review studies have shown that, in real-world financial risk applications, retaining the original imbalanced distribution better captures actual business logic and avoids a decline in accuracy caused by the introduction of synthetic noise [9].

From a practical view, this high precision is very valuable for investors. In the stock market, a false alarm happens when a healthy company is wrongly labeled as a risk. This mistake leads to high costs, such as wasted time and lost opportunities. Earlier studies on credit scoring also show that prioritizing precision usually saves more money than just chasing a high recall rate [10]. Since the model ensures that its warnings are likely to be correct, it provides high-conviction signals that users can trust.

DOI: 10.54254/2753-8818/2026.33745

5. Conclusion

This study has confirmed that XGBoost is an effective tool for financial early warning in the Chinese A-share market. By using raw and unbalanced data, this model can provide reliable signals and meet the needs of actual investment decisions. The analysis and classical theories have both confirmed that profitability indicators are the main drivers of financial risks.Although interpretable artificial intelligence has enhanced transparency, there are still some limitations. The current research relies on static reports and does not consider temporal changes. Future research should attempt to incorporate unstructured data, such as annual report texts, and combine it with macroeconomic variables. Transitioning from static checks to real-time monitoring will provide more powerful technical support for risk prevention in the capital market.

References

[1] Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 22, 59. https: //doi.org/10.2307/2490859
[2] Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https: //doi.org/10.1145/2939672.2939785
[3] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https: //doi.org/10.1016/j.inffus.2019.12.012
[4] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109. https: //doi.org/10.2307/2490395
[5] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https: //doi.org/10.1007/s10462-020-09896-5
[6] Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine Learning Explainability in Finance: An application to default Risk analysis. SSRN Electronic Journal. https: //doi.org/10.2139/ssrn.3435104
[7] Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems With Applications, 78, 225–241. https: //doi.org/10.1016/j.eswa.2017.02.017
[8] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232. https: //doi.org/10.1007/s13748-016-0094-0
[9] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2016). Learning from class-imbalanced data: Review of methods and applications. Expert Systems With Applications, 73, 220–239. https: //doi.org/10.1016/j.eswa.2016.12.035
[10] Brown, I., & Mues, C. (2011). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems With Applications, 39(3), 3446–3453. https: //doi.org/10.1016/j.eswa.2011.09.033