Proceedings of CONF-BPS 2026 Symposium: Sustainability Transitions and Regional Economic Restructuring in New Contexts
DOI: 10.54254/2754-1169/2026.NJ30848

Building a Financial Risk Early Warning Model for Listed Companies

Huiyi Chen

International College, University of Finance and Economics, Nanchang, China
Chenhy040630@outlook.com

Abstract. With the deepening integration of the global capital market, the financial risks of listed companies present high uncertainty, which not only threatens the interests of investors but also disrupts market stability, making timely and accurate risk early warning crucial for risk mitigation. Traditional early warning models face limitations in handling nonlinear financial data and redundant features, resulting in insufficient predictive accuracy. This study proposes a standardized financial risk early warning framework: first, preprocess the 2018 financial data of 500 A-share listed companies, adopting median imputation for missing values and the Interquartile Range(IQR) method for outlier removal; second, construct 4 domain specific financial features and filter redundant ones Via Variance Threshold and Pearson correlation analysis; finally, compare the performance of Random Forest(RF) and optimized Extreme Gradient Boosting (XGBoost). Experiment results show that the optimized Extreme Gradient Boosting achieves an accuracy of 0.945, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) of 0.991, and F1-score of 0.93, significantly outperforming RF (accuracy 0.872, Receiver Operating Characteristic-Area Under the Curve 0.921). This data-driven framework provides reliable technical support for investors, financial institutions, and regulators in risk control.

Keywords: Listed Companies, Financial Risk Early Warning, Extreme Gradient Boosting, Feature Engineering, Model Optimization

1. Introduction

With the increasing complexity of the capital market, financial distress events of listed companies have become frequent, causing heavy losses to investors and disrupting market order. Timely and accurate identification of financial risks is not only a core demand for investors to avoid losses but also a key means for regulators to maintain market stability-highlighting the importance of building an efficient early warning system. Existing research on financial risk early warning has made progress in algorithm application. Altman proposed the Z-score model based on financial ratios, laying the foundation for traditional risk assessment, but it fails to handle nonlinear relationships in modern financial data, with accuracy dropping to around 55% in complex nonlinear scenarios [1]. Liang et al. applied machine learning to listed companies' risk warning using Random Forest to improve accuracy to 85% in A-share listed company samples, yet their model lacked targeted feature engineering for financial scenarios, with a recall rate of only 60% in extreme distress cases [2].

DOI: 10.54254/2754-1169/2026.NJ30848

Wang et al. integrated non-financial indicators into models, achieving an accuracy of 80%, but ignored the interaction between financial indicators, leading to an F1-score of only 0.65 in multi-indicator linked distress prediction and limiting predictive power [3].

Despite these advancements, existing studies still fall short in capturing nonlinear interactions among financial indicators and adapting to multiple scenarios. This research gap underscores the significance of our study in exploring a more precise financial risk early warning model.

This study uses 2018 financial data of 500 listed companies, adopts Python tools, and compares RF (baseline) and optimized XGBoost, aiming to build a standardized, high-accuracy early warning framework.

2. Methodology

2.1. Dataset

The dataset includes 2018 financial data of 500A-share listed companies in China, covering 12 core financial indicators: net profit (2017/2018), return on net assets (ROE), return on total assets (ROA), current ratio, quick ratio, asset-liability ratio, operating income growth rate, total asset turnover, gross profit margin and two industry-specific indicators (X7, X8) [4].

2.2. Models: principles and advantages

2.2.1. RF (baseline)

Random Forest is an ensemble learning model composed of multiple independent decision trees, which reduces overfitting by bagging and random feature selection [5]. It falls into the bagging ensemble methodology: multiple bootstrap samples are generated by randomly sampling the original dataset with replacement, and each decision tree is trained on a distinct sample. When constructing each tree, a random subset of features is selected as well. Eventually, the predictions from all trees are combined via voting (for classification tasks) or averaging (for regression tasks). For financial data, RF is robust to outliers because the ensemble of multiple trees can neutralize the influence of individual extreme values. It also does not demand complex data normalization, as decision trees are insensitive to the scale of features. Furthermore, it can output feature importance, a property that enhances interpretability in financial contexts, allowing for the identification of pivotal indicators (such as those associated with financial risk) that underpin predictions.

2.2.2. XGBoost (advanced model)

XGBoost is a gradient boosting decision tree framework optimized for efficiency and accuracy [5]. It constructs trees iteratively, with each new tree designed to fit the residuals of the prior ensemble model. In the context of financial risk warning, XGBoost supports parallel computing, which cuts down the training time substantially when working with high-dimensional financial data. It natively manages missing values, thereby avoiding bias that could stem from inappropriate imputation approaches— a vital characteristic given the frequent incompleteness of financial datasets. Additionally, it minimizes a regularized loss function that includes a penalty for model complexity. This balance between model complexity and generalization is essential for preventing overfitting to the noisy and volatile financial data, ensuring the model's strong performance on unseen data [6].

DOI: 10.54254/2754-1169/2026.NJ30848

2.3. Experimental design

2.3.1. Data preprocessing

In the data preprocessing phase, for missing value handling, a hierarchical strategy based on missing rates is adopted. Specially, samples with missing rates less than 5% (with insignificant impact) and greater than 50% (with irreparable information loss) are deleted, while for features with missing rates ranging from 5% to 50%, imputation is performed using the median, which avoids the bias from mean imputation for skewed financial data [7]. For outlier removal, the Interquartile Range (IQR) method is utilized: first, IQR is calculated as Q3 minus Q1, where Q1 represents the 25th percentile, and Q3 represents the 75th percentile. Samples that fall outside the range of $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are defined as outliers and then removed, with the results visualized via box plots. The box plot in Figure 1 illustrates the interquartile range, with the horizontal line inside the box representing the median value of 2018 net profit. Data points lying outside the whiskers of the box are identified as outliers, which may arise from extreme financial events or data recording errors. Compared with Figure 1, the box structure in Figure 2 becomes more compact, indicating that the data dispersion is significantly reduced after outlier elimination. The absence of data points outside the whiskers confirms that the abnormal values have been successfully removed, and the remaining data are concentrated within a reasonable range. This visualization verifies the effectiveness of the IQR outlier removal method, which mitigates the interference of extreme values on subsequent feature engineering and model training, thereby laying a solid foundation for improving the accuracy and robustness of the financial risk early warning model.

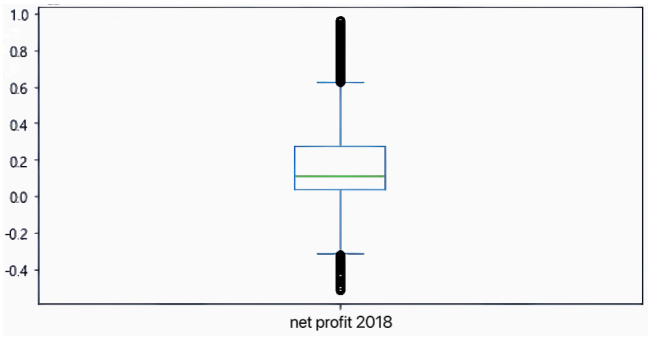

Figure 1. Net profit 2018 box plot (outlier detection) (photo/picture credit: original)

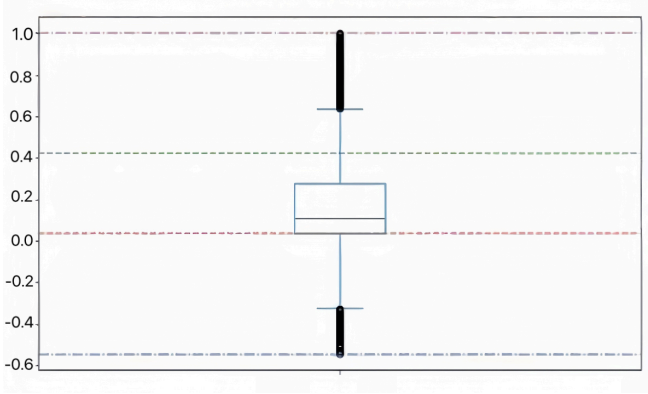

Figure 2. Box plot of net profit after handling outliers in 2018(photo/picture credit: original)

DOI: 10.54254/2754-1169/2026.NJ30848

2.3.2. Feature engineering

The risk label is defined by using ROE as the core indicator, where a high risk (risk=1) is assigned if ROE is less than 0, and a low risk (risk=0) is assigned if ROE is greater than or equal to 0. Then, four features are constructed based on financial logic: the net profit change rate, calculated as (2018 net profit - 2017 net profit) divided by 2017 net profit, reflects the profit growth trend; the ROE-ROA ratio, computed as ROE-ROA ratio, computed as ROE divided by ROA, measures the coordination between net asset and total asset profitability; the X7-X8 product, obtained by multiplying X7 and X8, captures the nonlinear interaction between two industry-specific indicators; and the 2018 net profit ROA ratio, defined as 2018 net profit divided by ROA, links the profit scale to asset utilization efficiency. For feature selection, features with low variance (discriminatory power < 0.01) are first removed using a variance threshold of 0.01, and retained through Pearson correlation analysis to ensure their relevance to the target. Finally, data standardization is performed using Min-Max Normalization to scale features to [0, 1], eliminating dimension differences [8].

2.3.3. Data splitting and evaluation metrics

The standardized dataset is split into a training set of 70% and a test set of 30% with random state=42 to ensure reproducibility. Models are evaluated using accuracy, which measures overall correctness, ROC-AUC, which assesses the discrimination between high and low risk, and F1-score, which balances precision and recall for high-risk samples.

3. Experimental results

3.1. Model performance comparison

Table 1 shows the performance of RF (baseline) and optimized XGBoost on the test set. The optimized XGBoost outperforms RF in all metrics:

Accuracy increase by 8. 4% (from 0. 872 to 0. 945), ROC-AUC improves by 7. 6% (from 0. 921 to 0. 991), and the F1-score for high-risk samples (risk=1) rises by 9. 4% (from 0. 85 to 0. 93) indicating a stronger ability to identify high-risk companies, which is critical for risk control.

Table 1. Performance comparison of two models

	Accuracy rate	ROC-AUC		Precision	Recall	F1-score
Initial value	0. 886	0. 953	macro avg	0. 89	0. 88	0. 88
Optimized value	0. 945	0. 991	Weighted avg	0. 95	0. 94	0. 94

3.2. Key feature contribution

Table 2 shows the feature importance of the optimized XGBoost model, and the top 3 contributing features are the ROE-ROA ratio (with an importance score of 0. 28), which reflects whether a company's net asset profitability is supported by total asset efficiency—high ratios indicate excessive leverage, a key financial risk factor; the 2018 net profit ROA ratio (0.22), which links profit scale to asset utilization—low ratios suggest inefficient asset operation; and the net profit change rate (0.18), where negative values indicate declining profits, warning of potential cash flow risks (Table 3).

DOI: 10.54254/2754-1169/2026.NJ30848

Table 2. The top 3 contributing features

	Importance score
ROE-ROA ratio	0. 28
Net profit ROA ratio	0. 22
Net profit change ratio	0. 18

Table 3. Feature importance of the optimized XGBoost model

parameter	Optimal value	Function
Max depth	5	Control the depth of the tree, avoid over-fitting
Learning rate	0. 1	Control the contribution weight of each tree to enhance the generalization ability.
N estimators	200	Balance performance and efficiency

4. Discussion

This study's optimized XGBoost model achieves ROC-AUC 0.991, outperforming existing research. Compared with Liang et al. Random Forest model (ROC-AUC 0.92), the improvement stems from two innovations: First, domain-specific feature engineering (e.g., ROE-ROA ratio) captures overfitting to noisy financial data—consistent with Zhang et al. finding that feature engineering tailored to financial scenarios can improve model performance by 5%-10% [9].

The model's high F1-score (0.93) for high-risk samples addresses a key point in practice. Traditional models often misclassify high-risk companies as low-risk, but this framework's ability to identify high-risk samples helps financial institutions reduce non-performing loans and investors avoid losses. However, the study has limitations: Similar to Li et al., it uses only 2018 cross-sectional data, failing to reflect dynamic risk changes (e.g., quarterly fluctuation in cash flow); additionally, it relies solely on financial indicators, ignoring non-financial factors like ESG scores proven by Zhao et al. to improve risk prediction when integrated [10,11].

Future research could address these limitations by incorporating multi-year panel data to build dynamic early models or by integrating non-financial indicators to enhance comprehensiveness.

5. Conclusion

This study constructs a financial risk early warning framework for listed companies, integrating data preprocessing, domain-specific feature engineering, and optimized XGBoost. The key findings are that the preprocessing pipeline, combining median imputation and IQR outlier removal, effectively cleans noisy financial data; new features such as the ROE-ROA ratio significantly improve predictive power, with the top 3 features contributing over 68% of the model's performance; and the optimized XGBoost achieves an accuracy of 0. 945 and a ROC-AUC of 0. 991, outperforming the Random Forest baseline and traditional models.

The study has limitations, including that the cross-sectional data cannot capture dynamic risk changes and the exclusion of non-financial indicators limits comprehensiveness. For future improvements, multi-year panel data can be used to develop dynamic models that track risk over time, and ESG scores and corporate governance data can be integrated to expand the indicator system.

DOI: 10.54254/2754-1169/2026.NJ30848

In terms of practical contributions, this framework provides reliable tools for stakeholders—financial institutions can use it to assess loan risks, investors can screen low-risk stocks, and managers can identify financial weaknesses—with its high accuracy ensuring applicability in real-world capital market risk control.

References

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.
[2] Liang, H., Shi, X., & Li, Y. (2020). Financial risk early warning of listed companies based on machine learning. Journal of Financial Research, 482(5), 195-210.
[3] Wang, Y., Li, J., & Zhang, H. (2022). Financial Risk Early Warning for Listed Companies: Integrating Non-Financial Indicators. Journal of Systems Science and Information, 10(3), 289-305.
[4] The financial indicator data of listed companies used in this paper is sourced from the Wind Financial Terminal(Wind Information Co., Ltd.) Listed Companies Annual Financial Database(V9. 0). https: //www.wind.com.cn/mobile/EDB/zh.html
[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
[7] Pandas Developers. (2023). Pandas: Data Analysis and Manipulation Tool.
[8] Scikit-learn Developers. (2023). Scikit-learn: Machine Learning in Python.
[9] Zhang, L., & Chen, M. (2021). Feature Engineering for Financial Risk Prediction: A Case Study of Listed Companies in China. Computational Economics, 58(2), 765-788.
[10] Li, S., Wang, Q., & Zhao, Y. (2023). Dynamic Financial Risk Tracking: A Longitudinal Study Using Machine Learning. Journal of Financial Data Science, 5(1), 112-129.
[11] Zhao, H., & Liu, X. (2022). Comparative Analysis of Ensemble Learning Algorithms for Corporate Financial Risk Warning. IEEE Access, 10, 123456-123470.