

Model Selection and Threshold Selection in Financial Distress Prediction: An Empirical Analysis Based on Listed Manufacturing Companies in China

Siqi Duan

School of Economics and Management, East China Normal University, Shanghai 200062, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: This paper analyses 3,837 manufacturing companies listed on the Shanghai and Shenzhen A-share markets between 2010 and 2024. It uses two consecutive years of losses as an indicator of financial distress and employs the T-2 offset forecasting method to avoid forward-looking bias. The data are divided chronologically into training and testing sets, and a comprehensive comparison is made of five commonly used early warning methods. Furthermore, the paper proposes a PAT dynamic threshold adjustment approach that accounts for time-series characteristics. The study concludes that there is limited scope for improvement in machine learning models; XGBoost only marginally improves the AUC relative to the baseline Logit model. If 0.5 is used as a fixed threshold, the early warnings issued by the model are largely ineffective, and non-threshold models are of little practical use. The threshold is the key determinant of the success of early warning; it is a more powerful indicator of cost issues than the model itself. Dynamic PAT thresholds can significantly reduce the cost of early warnings and markedly improve the identification of financial distress. From a practical perspective, financial risk early warning systems should focus on cost-oriented threshold setting rather than simply increasing model complexity.

Keywords: Financial distress prediction; Decision threshold; Model selection; Class imbalance; PAT dynamic threshold

Online publication: September 30, 2026

1. Introduction

Early warning of financial distress is a risk control measure in capital markets; it is of great practical significance, as it assists creditors in making credit arrangements and preventing debtor defaults, enables investors to gain a clearer understanding of the risks associated with the companies in question and select more favorable options, and facilitates regulators in monitoring the market in accordance with established procedures, minimizing risk losses and promoting more efficient use of capital. Data on listed manufacturing companies exhibit significant imbalances in terms of both categorical distribution and time-series benchmarks; however, this has not yet been sufficiently addressed in academic literature. From a practical perspective, analyzing the thresholds and criteria used in early warning systems generally yields better predictive results than simply

refining model algorithms.

The prediction of financial distress is a topic currently receiving significant attention within the domestic academic community, and a substantial body of literature has provided comprehensive analyses of relevant models and algorithms. Early research primarily employed logistic regression; Li ^[1] and Guo ^[2] developed early-warning models using data from listed manufacturing companies, thereby laying a solid methodological foundation for subsequent research. Machine learning is now gradually making inroads into this field of research: Wang ^[3] and Xiang ^[4] examined LASSO-Logit, XGBoost, and Random Forests as tools, systematically comparing the predictive capabilities of these various models; Song ^[5] and Zhang ^[6] further improved model design from the perspectives of feature selection, parameter setting, and even resampling, resulting in a marked improvement in early warning performance. Zou has reflected on traditional evaluation metrics relating to imbalanced samples and provided a more detailed explanation of the relevant evaluation criteria ^[7].

Currently, the academic community predominantly employs matching or resampling techniques to address data imbalance, conducting classification analyses based on predetermined thresholds. However, there has been little discussion regarding threshold optimization under the true distribution, nor has a quantitative comparison been made between threshold setting and model selection.

2. Research design

2.1. Sample and data

The sample used in this paper comprises 3,837 listed manufacturing companies on the Shanghai and Shenzhen A-share markets from 2010 to 2024; focusing on a single sector helps to avoid inconsistencies in financial benchmarks arising from the inclusion of multiple sectors. Following appropriate transformation of the initial data, a preliminary sample of 46,372 valid indicators was selected. This was then stratified and filtered based on the definition of financial distress using the complete data status for the T–2 period, resulting in a final set of 37,983 observations for modeling.

2.2. Dependent variable

The paper uses two consecutive years of negative net profit (D1) as the criterion for determining financial distress; this criterion corresponds to the basic financial trigger conditions set out when the Shanghai and Shenzhen Stock Exchanges established the Special Treatment (ST) system in 1998 ^[4]. To ensure the robustness of the empirical results, two commonly used metrics have been excluded here: the regulatory criterion D2, which introduced new revenue assessment standards following amendments to the listing regulations in 2021, resulting in a sharp drop in the distress rate from 3.63 percent in 2020 to 0.48 percent in 2021; this policy change caused a data break and an insufficient test sample; the ‘insolvency’ measure (D3) involved only approximately 30 test cases, rendering it statistically insignificant. As D1 takes into account net profits for both period T and period T–1, using period T–1 variables for forecasting would result in significant forward-looking errors; therefore, this paper employs period T–2 indicators to design a lagged forecast for period T ^[6].

2.3. Explanatory variables

The explanatory variables listed in this paper are based on Ohlson’s nine-variable method, with eight additional items incorporated on the basis of major domestic research, totaling 16 financial indicators. These provide a relatively comprehensive analysis of various types of enterprises (classified by scale, financial leverage,

liquidity, profit, cash, operations, development, and interest coverage). The empirical results indicate that the proportion of missing interest coverage ratios is 38.16 percent. Although this indicator is associated with financial distress during the same period, the situation during that period cannot be incorporated into the model because the ICR was not disclosed in the forecast year, and the loss incurred in that year represents a non-causal relationship between the two. Here, rather than excluding missing values, the method of imputing zeros and adding indicator variables is employed; this approach neither discards 29.4 percent of the observational data nor obscures the logic behind the missing values.

3. Research methods

3.1. Partitioning and modeling

The research discussed in this paper is based on a framework comprising time-series modeling, imbalanced data adaptation, and dynamic threshold optimization. With regard to the issue of information leakage, the paper employs a strict temporal division, designating the period from 2012 to 2020 as the training period, comprising 23,030 observations and 476 distress events, with a baseline ratio of 2.07 per cent; the period from 2021 to 2024 was designated as the test period, comprising 14,953 data points and 1,265 distress events, with a baseline rate of 8.46 per cent; at this point, a marked temporal drift of approximately 4.09 times was observed in the sample baseline rate (see Figure 1). The model categories selected comprised five forms: baseline Logit, PCA-Logit, LASSO-Logit, Random Forest and XGBoost. When processing variables using the PCA approach, principal component analysis was performed using only 14 financial ratios; 10 principal components were obtained by retaining 95 percent of the variance, whilst binary dummy variables were incorporated as they were, without any redundancy or weak information. Had PCA been applied to all variables, the model's AUC would have deteriorated. Regarding parameter tuning, the penalty level for LASSO-Logit is determined via internal 5-fold cross-validation, whilst the parameters for the other models are adjusted using a time-series cross-validation method on the training set; the test set is never involved in parameter tuning, thereby enabling a relatively reliable prediction of the model's out-of-sample performance.

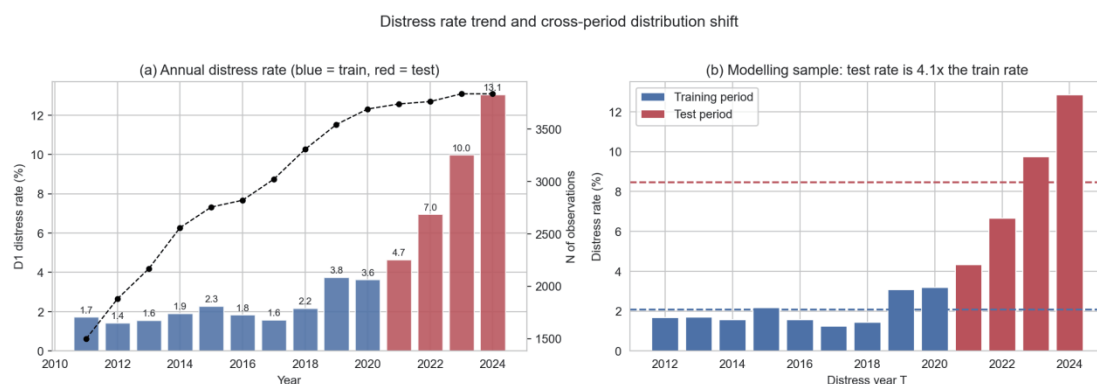


Figure 1. Annual trends in the rate of people in difficult circumstances

3.2. Unbalanced processing bandwidth and PAT dynamic threshold

Empirical analysis reveals that the use of class weighting during training significantly impairs the quality of the model's probability outputs: the weighted predicted probabilities are often biased, the Brier score deteriorates,

probability calibration fails to hold, and the resulting predictions are far inferior to those of a constant-threshold model. In contrast, threshold adjustment does not affect the probabilities obtained during training or their relative order, and no probability distortion occurs. To address the failure of static thresholds caused by benchmark rate drift, a dynamic threshold scheme of the PAT (Prior-Adjusted Threshold) type is employed here, treating the costs of false negatives and false positives in a 10:1 ratio. The paper performs a prior correction based on the likelihood ratio, defines the drift multiplier, and derives a closed-form solution for PAT:

As can be seen from the formula, when the baseline rate rises and λ , the threshold shifts downwards, thereby increasing the sensitivity of the early warning system. Taking the test period (training benchmark rate of 2.07 percent) as an example, between 2021 and 2024, λ rose from 1.56 to 5.12, whilst τ decreased year on year from 0.0601 to 0.0192, with a mean of 0.0381. This figure closely matches the empirically determined threshold of 0.0380 obtained through numerical optimization, thereby validating the correctness of the derivation.

4. Empirical results

4.1. Main results

The main results in this paper are uniformly reported using the unweighted model and the cost-sensitive threshold (PAT); the average thresholds for each model are shown in Table 1.

Table 1. Comparison of the performance of the five models on the test set (2021–2024)

Model	Average threshold	AUC-ROC	AUC-PR	Precision	Recall	Specificity	G-mean	F1	MCC	Brier
Logit	0.0385	0.8403	0.3485	0.2831	0.6775	0.8415	0.7550	0.3993	0.3594	0.0738
PCA-Logit	0.0348	0.8404	0.3468	0.2760	0.6767	0.8360	0.7521	0.3921	0.3519	0.0739
LASSO-Logit	0.0410	0.8401	0.3502	0.3046	0.6427	0.8644	0.7453	0.4133	0.3685	0.0739
Random Forest	0.0423	0.8740	0.4080	0.3052	0.7423	0.8438	0.7914	0.4325	0.4034	0.0720
XGBoost	0.0336	0.8772	0.4152	0.2857	0.8063	0.8137	0.8100	0.4219	0.4047	0.0731

When assessed in terms of ranking performance, the AUC-ROC values of 0.8401 to 0.8772 obtained by the five models are relatively consistent (see Figure 2). XGBoost performed best on the classification task, achieving an AUC of 0.8772; however, compared with the baseline Logit model, this represented an improvement of only 0.037, indicating a modest overall gain. For the Logit model itself, as well as the three linear models—PCA-Logit and LASSO-Logit—the AUC values largely converged, fluctuating around 0.840; that is, the improvement in ranking performance achieved by machine learning is essentially negligible compared to traditional methods. After LASSO-Logit identified the optimal penalty term using five-fold cross-validation, it retained 11 effective variables. At this point, the AUC was almost identical to that of the baseline Logit model, suggesting that, when financial ratios exhibit high multicollinearity, variable selection neither improves the model nor enhances predictive accuracy.

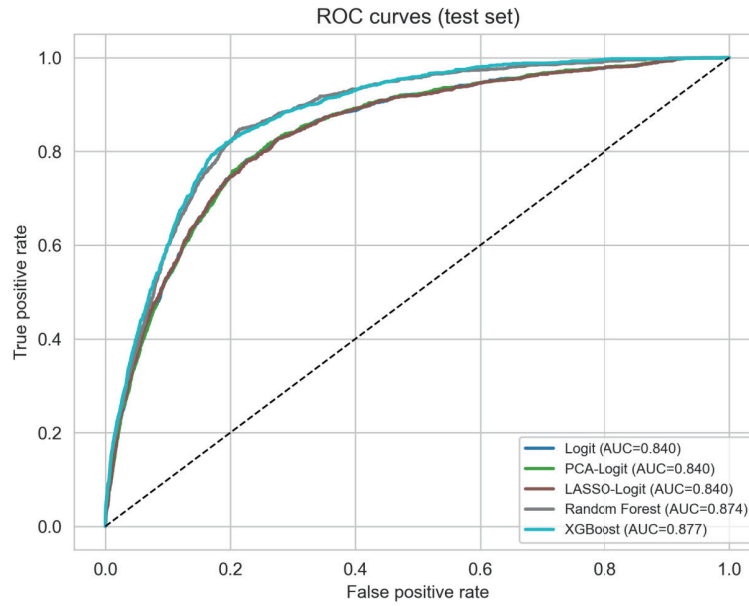


Figure 2. ROC curve

For PR curves where class imbalance is relatively easy to handle (see Figure 3), the gap between the models in question widens slightly. The average precision of XGBoost is 0.4152, whilst that of Logit is 0.3485.

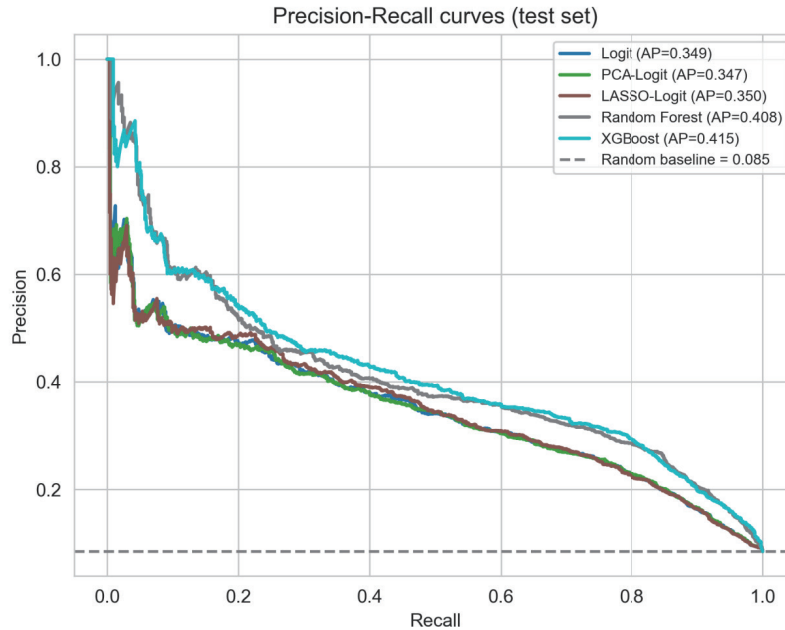


Figure 3. PR curve

4.2. The early warning effect of the PAT dynamic threshold

Based on the four categories of threshold rules described in Table 2, we examine the value of PAT's dynamic thresholds in each year in mitigating the failure of traditional early warning systems. The four types of rules

exhibit a clear performance gradient: whilst using 0.5 as the default threshold is easy to implement, it is of no practical significance; in this case, the total cost is 12,469, and the recall rate is only 1.50 percent. The static optimal threshold identified by OOF during the training period is slightly conservative due to fluctuations in the temporal baseline rate; the total cost is reduced to 9,221, with a maximum recall rate of 31.30 percent. The PAT dynamic threshold mentioned in the paper represents the optimal implementable solution; its average value of 0.0380 corresponds to a total cost as low as 5,872 and a recall rate as high as 70.91 percent. The PAT dynamic threshold effectively addresses the temporal drift in the baseline rate, compensating for the shortcomings of a fixed threshold, and serves as empirical evidence of the relationship between the threshold and the model.

Table 2. A comparison of the performance of four threshold rules (cost ratio 10:1)

Threshold rules	Average threshold	Total cost	Precision	Recall	Specificity
Rule 0:Default 0.5	0.500	12,469	0.401	0.0150	0.9993
Rule 1:Optimal OOF during the training period	0.083	9,221	0.429	0.3130	0.9612
Rule 2:PAT (Year-on-Year)	0.0380	5,872	0.291	0.7091	0.8399
Rule 3:Oracle (Optimal on the test set)	0.024	5,537	0.257	0.7935	0.7863

4.3. Thresholds and the relative importance of models: Evidence from variance decomposition

This paper will analyze the relative importance of the selected models and thresholds to the early warning results from the perspectives of range comparison and variance decomposition. An examination from the perspective of range shows that, if the threshold remains unchanged, the additional cost incurred by switching to the model is merely 308, whilst the cost fluctuation associated with a single model spanning the entire threshold grid reaches 7,113—the latter being 23.1 times that of the former; variance decomposition further illustrates the proportions between the two (see Figure 4).

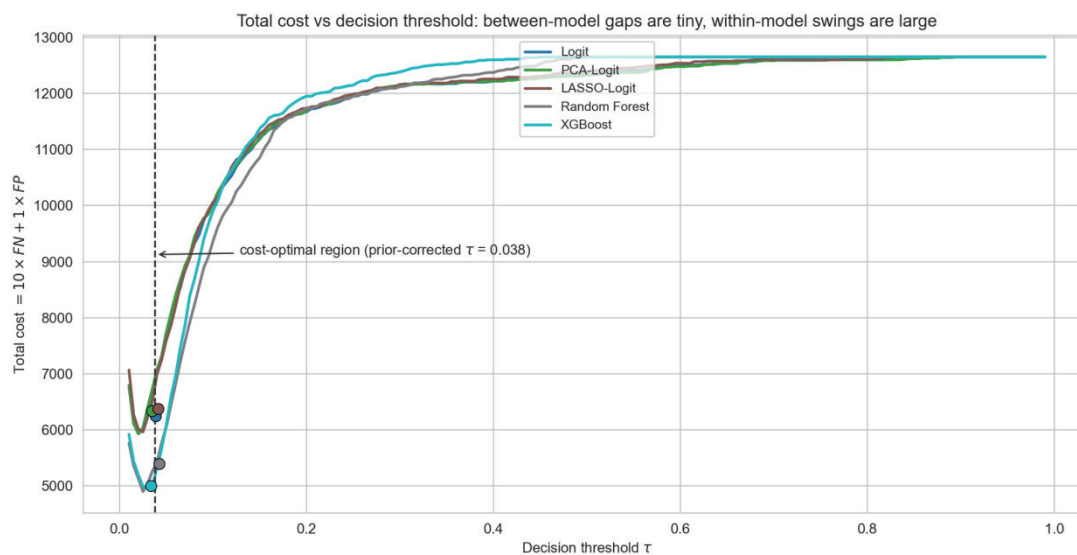


Figure 4. Total cost as a function of changes in the decision threshold (cost ratio of 10:1)

A variance decomposition based on the baseline scenario reveals that the choice of threshold accounts for 98.34 percent of the variation in total costs, whilst model selection accounts for only 0.06 percent. The results

presented here do not negate the existence of performance differences between models; however, inherent differences between models should not be mistaken for the sharp fluctuations in costs caused by inappropriate threshold selection. With regard to PAT, when calculated using the optimal threshold, the tree model yields cost savings of 13.6%–21.6% compared to the linear model, whilst model selection can increase the recall rate by approximately 12.88 percentage points. Threshold determination and model selection are not mutually exclusive steps; when forecasting financial distress, one should first set the threshold and define the objective function before considering the fine-tuning of the model.

5. Conclusion

The discussion takes the prediction of time-series financial distress as its starting point, conducting empirical tests within a framework of multi-model comparison and threshold optimization to draw key conclusions. From a machine learning perspective, the ranking prediction methods employed have inherent limitations; even the best-performing model, XGBoost, only achieves an AUC improvement of 0.037 compared to the Logit benchmark, indicating that complex algorithms cannot deliver significant performance breakthroughs across the board. Secondly, the insufficient model gain is primarily due to limitations inherent in the data itself, rather than insufficient model capacity; the observed learning curve remains fundamentally constrained by sample size, and there is an optimal threshold for the number of training iterations. With only 476 distress cases currently utilized, it is still difficult to drive the full development of complex models. Furthermore, regarding the critical significance of thresholds in early warning practice, the total cost variance attributable to threshold fine-tuning was 7,113–23.1 times that of model replacement. Thresholds account for 98.34 percent of the cost variance, whilst model selection accounts for only 0.06 percent. Threshold setting is a subsequent stage in model optimization rather than a substitute for it; at the optimal threshold, tree models save 13.6–21.6% in costs compared to linear models, and the selection of the optimal model still holds marginal value.

This study has two limitations: the sample used comprises only manufacturing enterprises, and therefore the conclusions may not be applicable to other sectors; the analysis did not include delisted companies in the sample, which may have resulted in an underestimation of the actual total market risk.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Li C, 2018, Construction of a Financial Risk Early Warning Model for Enterprises Based on Logistic Regression. *Statistics and Decision-Making*, 34(6): 185–188.
- [2] Guo E, 2020, Prediction of Financial Distress in Chinese A-share Listed Companies Based on the Logistic Model. *Business Accounting*, (10): 35–39.
- [3] Wang R, 2019, Prediction of Financial Distress in Listed Companies Based on Lasso-Logistic and XGBoost, thesis, Zhongnan University of Economics and Law.
- [4] Xiang Y, 2022, A Study on the Early Warning of Financial Distress in Listed Companies Based on Machine Learning, thesis, Southwest University.

- [5] Song Y, 2024, Prediction of Financial Distress in Information Technology Enterprises Based on Random Forest and SVM Algorithms. *Modern Computers*, 30(9): 43–50.
- [6] Zhang X, 2023, A Study on the Early Warning of Financial Distress in Listed Companies Based on Lasso-GWO-SVM, thesis, Shanghai Normal University.
- [7] Zou F, 2023, Prediction of Financial Distress in Small and Medium-sized Enterprises Based on Machine Learning. *Information Systems Engineering*, (12): 68–71.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.