

EARLY PREDICTION OF FINANCIAL DISTRESS RISK IN VIETNAMESE LISTED INDUSTRIAL FIRMS USING MACHINE LEARNING

Nguyen Thi Hai Hang

College of Foreign Economic Relations, Ho Chi Minh City

Email: nguyenhaihang2909@gmail.com

Received: 4 April 2026; Revised: 6 May 2026; Accepted: 14 May 2026

ABSTRACT

Early prediction of financial distress risk is important for providing timely warning signals to investors, creditors, and regulators. This study examines 91 Vietnamese listed industrial firms over the period 2015–2024 and compares a traditional Logistic Regression model with three machine learning algorithms: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). The dependent variable captures a firm's financial distress status in the following year, while the explanatory variables include conventional financial ratios, firm size, and lagged GDP growth. The empirical results for the test period show that all machine learning models outperform Logistic Regression in terms of overall predictive performance. RF achieves the highest Area Under the ROC Curve (AUC = 0.8741), followed by XGBoost (0.8578) and KNN (0.8030), while Logistic Regression records the lowest value (0.7232). At the 0.2 probability threshold, XGBoost provides the strongest distress detection, with the highest Recall of 76.92% and the highest F2-score of 71.68%, whereas RF offers a better balance between detecting distressed firms and controlling false alarms. Additional sensitivity analysis at the 0.3 probability threshold confirms the stability of the main findings. Furthermore, SHAP-based feature importance indicates that asset structure, profitability, leverage, and cash-flow capacity are the most influential predictors in the tree-based models. These findings support the use of machine learning, particularly RF and XGBoost, for early warning systems in Vietnamese listed industrial firms.

Keywords: Financial distress risk, Random Forest, XGBoost, K-nearest neighbors.

1. INTRODUCTION

The financial stability of listed firms plays a pivotal role in the sustainable development of Vietnam's stock market. In recent years, especially in the post-COVID-19 period with complex macroeconomic fluctuations, Vietnamese industrial enterprises have been facing unprecedented challenges. Given their capital-intensive nature and strong sensitivity to the business cycle, these firms become particularly vulnerable to financial distress. Therefore, the ability to accurately predict this risk is not only a matter of survival for managers in restructuring operations, but also an urgent requirement for investors and financial institutions in order to manage risk effectively.

Traditionally, financial risk prediction has relied mainly on statistical models. The seminal work of Altman [1] with the Z-Score model became a global benchmark thanks to its simple structure and relatively effective classification of bankrupt versus non-bankrupt firms in the data context of that period. Subsequently, Ohlson [2] developed the Logit model, which allows direct estimation of the probability of financial distress. Although these linear models

laid the foundation for the field of bankruptcy forecasting, they increasingly reveal limitations in the modern financial environment, where the relationships between financial ratios and corporate risk are nonlinear and are affected by the problem of imbalanced data, in which the number of healthy firms overwhelmingly exceeds that of distressed firms [3].

To overcome these limitations, the application of ML and Artificial Intelligence (AI) in financial forecasting has attracted growing attention. Many recent studies indicate that ML algorithms, especially ensemble techniques such as RF and XGBoost, can better capture nonlinear patterns, handle large and imbalanced datasets more effectively, and thereby significantly improve predictive accuracy compared with traditional statistical models [4], [5].

In the Vietnamese context, empirical studies that directly compare machine learning models with traditional approaches for specific industries remain relatively limited, particularly those using long panel datasets covering the post-COVID-19 period. This study is designed to address this gap by applying and comparing four models for predicting financial distress risk: Logistic Regression, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). Accordingly, the study makes several contributions. First, it provides additional empirical evidence on the early prediction of financial distress risk for Vietnamese listed industrial firms, a group of firms that is highly sensitive to business-cycle fluctuations and macro-financial conditions. Second, the study uses a dataset covering the period 2015-2024, which spans the pre-, during-, and post-COVID-19 periods., thereby better capturing changes in the business environment and corporate financial risk under the new market conditions. Third, the study defines the target variable in a forward-looking manner by predicting firms' risk status in the following year, making the results more consistent with the objective of early warning rather than merely identifying contemporaneous risk. Fourth, the study directly compares the traditional model with machine learning models using the same dataset and the same time-based data-splitting strategy, thereby providing a more objective basis for assessing the effectiveness of machine learning in financial distress prediction.

2. LITERATURE REVIEW ON FINANCIAL DISTRESS PREDICTION

2.1. Theoretical foundations of financial distress prediction

Financial distress is commonly understood as a gradual deterioration in a firm's financial health, starting from cash-flow difficulties and potentially leading to insolvency or bankruptcy if no timely intervention is made [3]. Recent studies argue that financial distress is not limited to legal bankruptcy events, but also encompasses intermediate states such as accumulated losses, covenant violations, being placed under warning or trading restrictions, and even delisting [4].

From a theoretical perspective, early prediction of financial distress is associated with two main theoretical frameworks. First, under Signaling Theory, financial ratios and risk indicators are signals that managers send to the market; an effective prediction model helps decode these signals, thereby reducing information asymmetry between firms and external stakeholders such as investors and creditors. Second, Agency Theory emphasizes conflicts of interest between managers and shareholders or creditors. An early-warning system functions as a disciplinary mechanism that constrains excessive risk-taking behavior or the concealment of unfavorable information. And in emerging markets, where legal bankruptcy data are still limited, many studies propose using alternative criteria to identify financial distress, such as low Z-Score, accumulated losses, or being placed under warning or control status [6], [3].

2.2. Studies using traditional financial distress prediction models

Research on bankruptcy or financial distress prediction began with traditional statistical models. Beaver [7] used univariate analysis to assess the classification power of individual

financial ratios. Altman [1] was the first to construct a multivariate model, the Z-Score, based on Multivariate Discriminant Analysis (MDA), combining five financial ratios to classify bankrupt and non-bankrupt firms. Thanks to its simple structure, ease of calculation, and interpretability, the Z-Score has been widely used as a benchmark tool for corporate risk screening in many markets [8].

However, MDA requires relatively strict assumptions regarding normal distribution and homogeneity of covariance matrices across firm groups. To overcome these limitations, Ohlson [2] proposed the Logit model to directly estimate the probability of financial distress. Logit is more flexible in handling explanatory variables and produces probability-based outputs that are convenient for risk management, thus gradually becoming a second benchmark alongside the Z-Score.

Nevertheless, both MDA and Logit are linear models, while in reality the relationships between financial ratios, macroeconomic conditions, and firm risk are often nonlinear with complex interactions [5]. Empirical data on financial distress are also typically highly imbalanced, with healthy firms greatly outnumbering distressed ones. This causes traditional models to be biased toward predicting the safe group and to miss many risky cases [3], [4]. Another line of research modifies the Z-Score structure or changes the set of financial variables to construct new Z-Score variants better suited to specific countries, industries, or study periods [8], [3]. However, these studies show that predictive performance still depends heavily on the characteristics of the sample data and does not fully resolve fundamental issues such as nonlinearity and data imbalance. This opens the need for more flexible models that can better capture complex data structures.

2.3. Studies using intelligent financial distress prediction models

The development of AI and ML has generated a new wave of research on financial distress prediction. The systematic review of Dasilas and Rigani [4], covering 207 empirical studies for the period 2012–2023, shows a clear shift from single traditional statistical models to ML algorithms and ensemble models. Approaches such as Support Vector Machines, Neural Networks, RF, Gradient Boosting, XGBoost and hybrid architectures are increasingly popular; abundant empirical evidence indicates that these models deliver higher accuracy and stability than Z-Score and Logit, especially in complex and imbalanced datasets.

Among them, RF is a bagging algorithm that builds multiple independent decision trees on bootstrap samples and aggregates their predictions. This mechanism helps reduce variance and mitigates the overfitting problem commonly observed in single decision trees. Studies such as Mai et al. [9] show that RF predicts corporate bankruptcy better than Logit in terms of both overall accuracy and the ability to identify distressed firms. Narvekar and Guha [5], when applying XGBoost to data on manufacturing firms in Italy, conclude that this model is particularly effective in handling imbalanced data and capturing nonlinear relationships among financial ratios.

In Vietnam, Truong and Le [6] are among the first to directly compare traditional models with ML. Using a multi-industry sample of 846 observations on listed firms over 2017–2019, the authors employ Logit and four ML algorithms (RF, XGBoost, KNN, and Naïve Bayes). The results show that RF and XGBoost outperform Logistic Regression in accurately classifying financially distressed firms, while Logistic Regression is still maintained as a reference model due to its interpretability. However, this study only covers the pre-COVID-19 period, relies on a relatively short observation window, and does not provide in-depth analysis for specific industries.

3. DATA AND RESEARCH METHODS

3.1. Data

Table 1. Independent variables in the model

Type	Variable name	Abbr.	Formula
Profitability/ Cash-flow coverage	Cash flow to total debt	CF_TD	$(\text{Net income after tax} + \text{Depreciation of fixed assets}) \div \text{Total liabilities}$
Size	Firm size	SIZE	Ln (Total assets)
Indebtedness	Debt ratio	DR	$\text{Total liabilities} \div \text{Total capital (total sources of funds)}$
Structure	Asset structure	AS	$\text{Current assets} \div \text{Total assets}$
Profitability	Return on assets	ROA	$\text{Net income after tax} \div \text{Total assets}$
Liquidity	Current ratio	CR	$\text{Current assets} \div \text{Current liabilities}$
Macroeconomic	Lagged GDP growth rate	LGDP	$(\text{GDP}_t - \text{GDP}_{t-1}) \div \text{GDP}_{t-1}$, using national real GDP

The literature on financial distress highlights the central role of traditional groups of financial ratios, as well as firm size. Altman [1], Ohlson [2], and Beaver [7] all show that indicators of profitability (ROA), liquidity (Current ratio), asset structure, and financial leverage are important explanatory variables for bankruptcy probability. In addition, firm size (SIZE), measured as the natural logarithm of total assets, is also considered a relevant predictor of financial risk, as larger firms are generally better positioned to access external financing, diversify their operations, and withstand adverse economic shocks [2]. Tinoco and Wilson [10] provide further evidence that leverage and liquidity ratios are statistically significant in distress prediction models for the UK market, while Rahman and Zhu [3] emphasize the role of cash flow relative to debt in the context of Chinese construction firms. At the macro level, GDP growth is widely used as a control variable reflecting overall economic conditions, which may reduce or increase financial distress pressure depending on the business cycle phase [4], [10]. On this basis, the present study chooses CF_TD, ROA, CR, DR, AS, SIZE, and LGDP as the main explanatory variables for the financial distress risk of listed industrial firms in Vietnam.

The sample focuses on 91 listed industrial firms in Vietnam. Financial statement data are collected over a 10-year period from 2015 to 2024, capturing multiple economic cycles, including the pre-, during- and post-COVID-19 stages. The dependent variable of the model is RISK_LEAD, a binary variable representing the firm's risk status in the following year. Following the criteria used in empirical research for Vietnam [6], this variable takes the value 1 if the firm is subject to compulsory delisting, is under warning, has accumulated losses, or has a Z-score below 1.81; and 0 if the firm is in a normal operating state.

Regarding the data-splitting strategy, instead of using a fixed random split, the study adopts a time-based split to better mimic a real forecasting setting and avoid looking ahead to future information. Specifically, because the dependent variable is defined for the following year, observations from 2015 to 2021 are used for model estimation, while observations from 2022 to 2023 are reserved for out-of-sample testing, corresponding respectively to the risk periods 2016–2022 and 2023–2024. To address the class imbalance problem, where the number of distressed firms is substantially smaller than the number of healthy firms, the study adjusts the relative weights of the two classes directly during model estimation.

3.2. Research methods

To test the hypothesis on the superiority of ML, the study builds and compares the performance of four classification models representing two different approaches. The benchmark linear, traditional model is Logit. The ML group consists of RF, XGBoost and KNN. All models are estimated using the same set of explanatory variables and the same time-based data split to ensure comparability.

Specifically, RF is constructed with 500 decision trees, using bootstrap sampling and random feature selection at each split to reduce variance and improve predictive stability. XGBoost is specified with a relatively small learning rate, limited tree depth, and class-weight adjustment so that the model pays greater attention to distressed firms under imbalanced data. KNN is implemented with a small number of neighbors and distance-based weighting, allowing the model to rely directly on local data structure in the space of financial ratios. Model performance is evaluated on the out-of-sample test set consisting of observations from 2022 to 2023, corresponding to the risk period 2023–2024. First, the study uses the area under the ROC curve (AUC) to assess the overall classification ability of each model. Next, at the 0.2 probability threshold, the study uses the confusion matrix to compute Type I error, Type II error, Recall, Precision, and F2-score. This approach makes it possible to evaluate not only overall predictive performance but also the trade-off between minimizing missed distressed firms and controlling false alarms in a financial risk management setting.

4. RESULTS AND DISCUSSION

The following table reports the descriptive statistics of the explanatory variables used in the model.

Table 2. Descriptive statistics of the variables

<i>Variable</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>Min</i>	<i>p25</i>	<i>p50</i>	<i>p75</i>	<i>Max</i>
CF_TD	910	1.4132	2.6770	-0.0011	0.2457	0.5959	1.4245	19.2868
DR	910	0.4567	0.2130	0.0027	0.2969	0.4634	0.6264	1.2945
AS	910	0.5900	0.2232	0.0226	0.4559	0.6232	0.7655	0.9833
ROA	910	0.0686	0.0782	-0.4673	0.0218	0.0530	0.1001	0.4503
CR	910	2.9976	5.2774	0.4427	1.1945	1.6110	2.4827	38.8730
SIZE	910	6.5019	1.4637	4.8271	5.8923	6.2593	6.6197	14.3512
LGDP	910	0.0893	0.0829	0.0291	0.0598	0.0675	0.0708	0.3345

Table 2 reports the descriptive statistics of the variables used in the model. All variables have 910 observations, indicating a balanced sample across firms and years. The financial variables display markedly different levels of dispersion. The cash flow to total debt ratio (CF_TD) has a mean of 1.4132 and a relatively large standard deviation of 2.6770, with values ranging from -0.0011 to 19.2868, suggesting substantial heterogeneity in firms' debt-servicing capacity from internal cash generation. The debt ratio (DR) has an average of 0.4567 and a median of 0.4634, with an interquartile range from 0.2969 to 0.6264, indicating that leverage is moderately concentrated around the middle range of the distribution. The asset structure variable (AS) has a mean of 0.5900, implying that current assets account for about 59% of total assets on average, although the wide range from 0.0226 to 0.9833 points to considerable variation in asset composition across firms.

Regarding firm performance and liquidity, the average return on assets (ROA) is 0.0686, while its minimum and maximum values of -0.4673 and 0.4503 reveal substantial differences

in profitability conditions among listed industrial firms. The current ratio (CR) has a mean of 2.9976 and a notably high standard deviation of 5.2774, with a maximum value of 38.8730, indicating that liquidity conditions vary widely and that some firms hold exceptionally large short-term asset buffers relative to current liabilities. Firm size (SIZE), measured as the natural logarithm of total assets, records a mean of 6.5019 and ranges from 4.8271 to 14.3512, reflecting significant differences in scale among firms in the sample. Finally, the lagged GDP growth variable (LGDP) has a mean of 0.0893 and a median of 0.0675, with a maximum of 0.3345, suggesting some right-skewness in the distribution of macroeconomic growth conditions over the study period. Overall, the descriptive statistics indicate substantial cross-sectional and intertemporal heterogeneity, providing a suitable empirical basis for financial distress prediction models in Vietnam's listed industrial sector.

To objectively assess the out-of-sample predictive performance of the models, the study uses a test set consisting of 182 observations corresponding to the risk period 2023–2024. As a first step, the AUC is examined. Table 3 below reports the AUC values, standard errors and 95% confidence intervals for the four models.

Table 3. The AUC results on the test set (TEST, N = 182)

<i>Model</i>	<i>AUC</i>	<i>Std. Err.</i>	<i>95% CI – low</i>	<i>95% CI – high</i>
Logit	0.7333	0.039	0.6556	0.8064
RF	0.8901	0.0241	0.8397	0.9326
XGB	0.8822	0.0243	0.8323	0.9251
KNN	0.7972	0.0368	0.7236	0.8678

Table 3 indicates that the traditional linear model, Logit, achieves an AUC of 0.7333, suggesting only a moderate ability to distinguish between distressed and non-distressed firms. In contrast, all three machine learning models perform substantially better, with AUC values well above 0.79. Among them, Random Forest (RF) delivers the highest AUC at 0.8901, indicating the strongest overall discriminatory power on the test set. XGBoost (XGB) follows closely with an AUC of 0.8822, further confirming the effectiveness of tree-based machine learning algorithms in capturing nonlinear relationships and complex interactions among financial indicators. KNN also improves meaningfully over Logit, with an AUC of 0.7972, although its performance remains below that of RF and XGB. In addition, the reported standard errors are relatively small, and the 95% confidence intervals are reasonably narrow for all models, especially RF and XGB, suggesting that the machine learning models not only achieve superior classification performance but also exhibit relatively strong stability in out-of-sample prediction.

Next, the study visually examines the classification outcomes of each model through the confusion matrices at the 0.2 probability threshold. The results are presented as follows:

Confusion Matrices on the Test Set

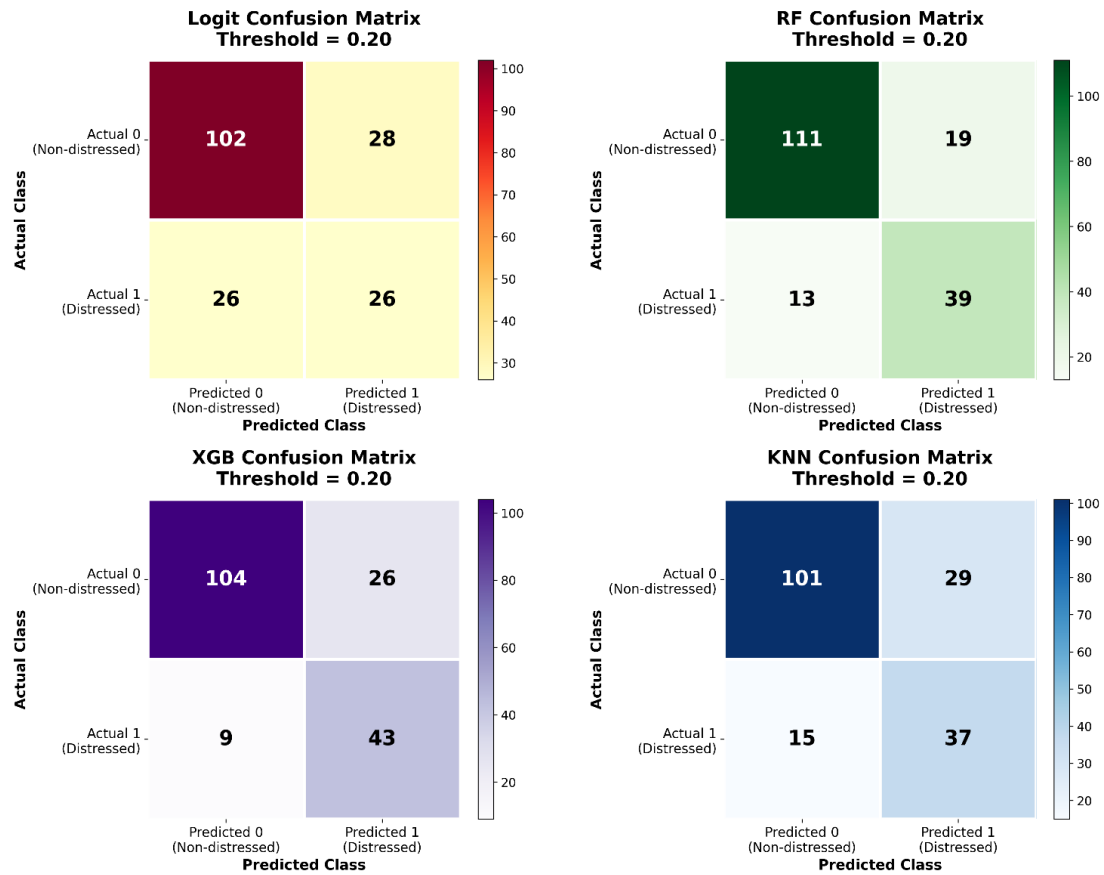


Figure1. Confusion matrices of Logit, RF, XGBoost, and KNN on the test set at the 0.2 probability threshold

Observations from the four confusion matrices reveal fairly clear differences across models in the trade-off between detecting distressed firms and generating false alarms for healthy firms. Visually, XGBoost appears to identify the largest number of distressed firms, while Random Forest shows a relatively better balance between correctly detected cases and false alarms. KNN also improves the detection of distressed firms compared with Logit, although its classification performance remains less balanced than that of the two tree-based models. These differences suggest that, beyond the overall classification ability measured by AUC, the practical usefulness of each model also depends on the relative priority placed on minimizing missed cases versus controlling false positives..

To quantify these observations more clearly, the study further summarizes the classification performance at the 0.2 probability threshold, including Type I error, Type II error, Recall, Precision, and F2-score. The detailed results are reported in Table 4 below.

Table 4. Classification metrics on the test set at the 0.2 probability threshold

<i>Model</i>	<i>Type I error</i>	<i>Type II error</i>	<i>Recall</i>	<i>Precision</i>	<i>F2-score</i>
LR	21.54	50	50	48.15	49.62
RF	14.62	25	75	67.24	73.31
XGB	20	17.31	82.69	62.32	77.62
KNN	22.31	28.85	71.15	56.06	67.52

Table 4 shows clear differences across the models when classification performance is evaluated at the 0.2 probability threshold. LR performs the worst overall. Although its Type I error is 21.54%, its Type II error reaches 50.00%, resulting in a Recall of only 50.00% and a Precision of 48.15%. This indicates that the linear benchmark model misses a substantial proportion of truly distressed firms and is therefore less suitable for early warning purposes. Among the machine learning models, XGBoost provides the strongest ability to identify distressed firms, with the highest Recall (82.69%) and the lowest Type II error (17.31%). It also achieves the highest F2-score (77.62%), confirming its superiority when the priority is to reduce missed distress cases. RF ranks second in terms of distress detection, with a Recall of 75.00% and a Type II error of 25.00%, but it performs best in controlling false alarms, as reflected in the lowest Type I error (14.62%) and the highest Precision (67.24%). KNN also outperforms LR, with a Recall of 71.15% and a Type II error of 28.85%, but it records the highest Type I error (22.31%) among the machine learning models and a lower Precision than both RF and XGBoost.

When evaluated by the F2-score, which places greater weight on Recall than on Precision, the ranking of the models continues to favor the machine learning group. Specifically, XGBoost achieves the highest F2-score (77.62%), followed by RF (73.31%) and KNN (67.52%), while LR reaches only 49.62%. These results suggest that if the main objective is to detect distressed firms as early as possible while tolerating a certain degree of false alarms, XGBoost stands out as the strongest model. By contrast, if decision-makers prefer a more balanced trade-off between distress detection and the reliability of warning signals, RF may be the more suitable choice.

The results of this study are broadly consistent with the trend highlighted in the review by Dasilas and Rigani (2024), which shows that machine learning models generally outperform the linear Logit model in financial distress prediction. For the sample of listed industrial firms over the period 2015–2024, all three machine learning models produce higher AUC values than Logit. Among them, RF achieves the highest AUC (0.8901), followed closely by XGBoost (0.8822) and then KNN (0.7972), while Logit reaches only 0.7333. These findings further support the view that machine learning algorithms, especially tree-based models, are better able to capture nonlinear relationships and complex interactions between financial indicators and distress risk. This pattern is also in line with earlier studies such as Mai et al. (2019), Narvekar and Guha, and Truong and Le (2023), which generally report superior classification performance of machine learning models over Logit in predicting corporate financial distress.

However, when performance is evaluated more closely at the 0.2 probability threshold, the results indicate that the optimal model is not the same under all criteria. XGBoost provides the strongest ability to identify distressed firms, with the highest Recall (82.69%), the lowest Type II error (17.31%), and the highest F2-score (77.62%), highlighting its advantage when the primary objective is to minimize missed distressed firms. Meanwhile, RF achieves the highest Precision (67.24%) and the lowest Type I error (14.62%), while still maintaining a relatively high Recall of 75.00%, indicating a better balance between correctly detecting distressed firms and controlling false alarms. KNN still performs better than LR, but remains

below the two tree-based models in terms of both AUC and classification performance at the selected threshold. Overall, these findings both reinforce and extend the existing empirical evidence. They suggest that machine learning is an appropriate approach for building early warning systems for financial distress, but that the choice of the optimal model should depend on the specific risk management priority between minimizing missed cases and accepting a certain level of false alarms.

To examine whether the classification results are highly dependent on the 0.2 probability threshold, the study further conducts a sensitivity analysis by increasing the probability threshold to 0.3. The classification results are presented in Table 5 below.

Table 5. Classification metrics on the test set at the 0.3 probability threshold

<i>Model</i>	<i>Type I error</i>	<i>Type II error</i>	<i>Recall</i>	<i>Precision</i>	<i>F2-score</i>
LR	18.46	51.92	48.08	51.02	48.64
RF	10.00	44.23	55.77	69.05	58.00
XGB	15.38	25.00	75.00	66.10	73.03
KNN	16.15	50.00	50.00	55.32	50.98

The results in Table 5 indicate that the main conclusions of the study remain stable. Specifically, XGBoost continues to provide the strongest ability to detect financially distressed firms, with a Recall of 75.00%, the lowest Type II error of 25.00%, and the highest F2-score of 73.03%. Meanwhile, Random Forest achieves the highest Precision of 69.05% and the lowest Type I error of 10.00%, indicating its advantage in controlling false alarms. Compared with the 0.2 threshold, increasing the threshold to 0.3 changes the trade-off between Recall and Precision, but does not alter the core conclusion that ensemble tree-based models, particularly XGBoost and Random Forest, remain more suitable than Logit and KNN for the early prediction of financial distress risk.

A clearer explanation for the empirical results can be derived from the characteristics of corporate financial data and the learning mechanism of each model. First, Logit, as a linear model, assumes that the effects of financial variables on the probability of financial distress follow a relatively stable and monotonic relationship. However, in practice, financial distress risk does not necessarily increase in a linear manner. For example, the debt ratio may not generate substantial risk when it remains at a moderate level, but once it exceeds a certain threshold, the pressure of interest and principal payments may cause the probability of financial distress to rise sharply. Similarly, the effects of liquidity, cash flow, or profitability often depend simultaneously on asset structure and macroeconomic conditions. Therefore, nonlinear relationships and interaction effects among financial indicators may weaken the predictive performance of Logit.

By contrast, Random Forest and XGBoost are decision tree-based models that can automatically split the data at multiple thresholds and capture complex interactions among variables without imposing a predefined linear functional form. Random Forest improves predictive stability by combining multiple decision trees constructed from different bootstrap samples, thereby reducing variance and limiting the risk of overfitting. In this study, this mechanism is clearly reflected in the fact that RF achieves the highest AUC among the four models, at 0.8901, with a relatively narrow 95% confidence interval ranging from 0.8397 to 0.9326. This suggests that RF not only provides the strongest overall discriminatory ability between distressed and non-distressed firms, but also maintains a relatively high level of stability on the test set.

XGBoost, on the other hand, uses a boosting mechanism in which subsequent trees focus on correcting the errors made by previous trees. This gives the model an advantage in

identifying difficult-to-detect distressed cases, especially under class-imbalanced data. In this study, this advantage is clearly reflected at the 0.2 probability threshold, where XGBoost achieves the highest Recall of 82.69%, the lowest Type II error of 17.31%, and the highest F2-score of 77.62%. This result indicates that XGBoost is particularly suitable when the priority is to minimize missed cases of firms with financial distress risk.

Although KNN also outperforms Logit, it depends heavily on the distance between observations and is therefore more sensitive to outliers, skewed distributions, and differences in variable scales. This is evident in the dataset used in this study, where some variables exhibit substantial dispersion, such as CF_TD ranging from -0.0011 to 19.2868 and CR ranging from 0.4427 to 38.8730. Therefore, although KNN achieves an AUC of 0.7972 and a Recall of 71.15%, both higher than those of Logit, it remains less stable than RF and XGBoost, as reflected in its lower Precision and the highest Type I error among the machine learning models, at 22.31%.

To enhance the interpretability of the machine learning models, the study uses SHAP-based feature importance to evaluate the contribution of each input variable in RF and XGBoost. The results are presented in Figure 2 below.

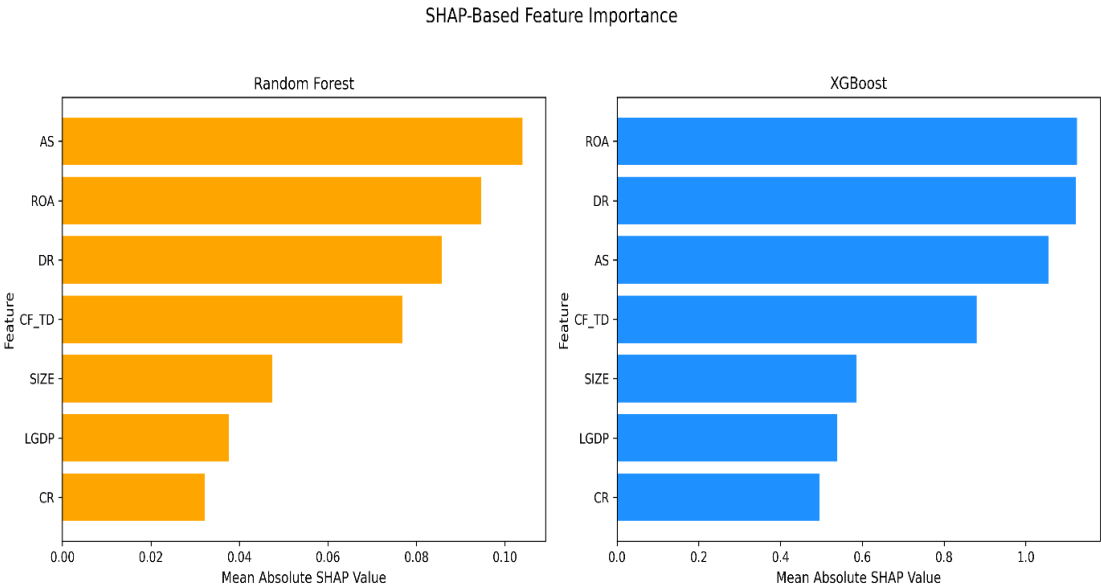


Figure 2. SHAP-based feature importance of Random Forest and XGBoost

The results in Figure 2 show that financial variables do not contribute equally to the prediction of financial distress risk. For the Random Forest model, the most important variables are AS, ROA, DR, and CF_TD, indicating that asset structure, profitability, leverage, and cash-flow capacity are key factors in distinguishing between distressed and non-distressed firms. For the XGBoost model, ROA and DR are the two most important variables, followed by AS and CF_TD. This result is consistent with financial reasoning, as lower profitability, higher leverage, less flexible asset structure, and weaker cash-flow capacity tend to increase the probability of financial distress. In addition, both models assign lower importance to CR, LGDP, and SIZE, suggesting that, in this study, short-term liquidity, macroeconomic growth conditions, and firm size play a more supplementary role compared with the above-mentioned financial indicators. Overall, the SHAP-based results improve the transparency of Random Forest and XGBoost by providing additional information on the relative importance of financial factors in predicting corporate financial distress risk.

5. CONCLUSION

Predicting financial distress is a key issue closely linked to a firm's solvency and long-term survival. For creditors, regulators, and investors, the early identification of firms at risk of financial distress helps support timely credit policy design, debt restructuring, risk monitoring, and more prudent investment decisions. Traditional statistical models such as the Z-Score and Logit have been widely used for decades. However, they rely on relatively strong assumptions about data structure and often face limitations when the relationships between financial ratios and firm risk are nonlinear, the data are imbalanced, and multiple complex interactions exist. The development of data science and machine learning has opened up a new generation of forecasting models that are more flexible and able to learn directly from historical data.

Based on a dataset of 91 listed industrial firms over the period 2015–2024, this study compares four financial distress prediction models: Logit, RF, XGBoost, and KNN, with the dependent variable reflecting the firm's risk status in the following year. The empirical results show that all three machine learning models achieve substantially higher AUC values than Logit. Among them, RF records the highest AUC (0.8901), followed closely by XGBoost (0.8822) and then KNN (0.7972), while Logit reaches only 0.7333. When the models are further evaluated on the test set at the 0.2 threshold, XGBoost shows the strongest ability to identify distressed firms, with a Recall of 82.69%, a Type II error of 17.31%, and an F2-score of 77.62%, indicating a clearer advantage when the objective is to minimize missed distressed firms. Meanwhile, RF achieves the highest Precision (67.24%) and the lowest Type I error (14.62%), while maintaining a Recall of 75.00%, reflecting a better balance between risk detection and false-alarm control. KNN also performs better than Logit, but overall remains below the two tree-based models. These findings further support earlier studies showing that machine learning algorithms, especially RF and XGBoost, have clear advantages over linear statistical models in financial distress prediction.

In terms of practical implications, the findings indicate that model selection should be aligned with specific risk-management objectives. XGBoost is more suitable in situations where the priority is to avoid missing firms with financial distress risk, such as in the early warning systems of banks, securities companies, or market supervisory agencies. With high Recall and low Type II error, XGBoost can identify a larger proportion of truly distressed firms, thereby helping creditors and investors proactively review credit profiles, reassess investment portfolios, request additional information, or implement preventive measures before the risk becomes more severe. By contrast, Random Forest is more appropriate when a balance between risk detection and false-alarm control is required. With higher Precision and lower Type I error, Random Forest can reduce the number of healthy firms incorrectly classified as risky, thereby saving resources for inspection, appraisal, and monitoring.

In addition, for banks, machine learning models can be integrated into early warning systems to support borrower screening, credit limit reassessment, adjustment of lending conditions, or proactive debt restructuring when the level of risk increases. For investors, the predicted results can be used as a supplementary screening tool to identify industrial firms showing signs of deteriorating financial health, thereby supporting portfolio allocation and investment risk management. For regulators and stock exchanges, early warning models can contribute to strengthening the monitoring of firms that may be subject to warning, control, or delisting, thereby enhancing market transparency and stability.

Although the study obtains several noteworthy findings, it still has certain limitations. The explanatory variables are mainly based on accounting-based financial indicators and GDP growth, and therefore do not fully capture market signals, corporate governance characteristics, and other non-financial factors. Future research may extend the dataset across

industries and incorporate market-based variables as well as governance variables such as ownership structure, board size, or the proportion of independent directors.

REFERENCES

- [1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968, doi: <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- [2] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, 1980, doi: <https://doi.org/10.2307/2490395>
- [3] M. J. Rahman and H. Zhu, "Predicting financial distress using machine learning approaches: Evidence from China," *Journal of Contemporary Accounting & Economics*, vol. 20, no. 1, Art. no. 100403, 2024, doi: <https://doi.org/10.1016/j.jcae.2024.100403>
- [4] A. Dasilas and E. Rigani, "Machine learning techniques in bankruptcy prediction: A systematic literature review," *Expert Systems with Applications*, vol. 255, Art. no. 124761, 2024, doi: <https://doi.org/10.1016/j.eswa.2024.124761>
- [5] A. Narvekar and D. Guha, "Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession," *Data Science in Finance and Economics*, vol. 1, no. 2, pp. 180–195, 2021, doi: <https://doi.org/10.3934/DSFE.2021010>
- [6] T. T. D. Truong and H. T. Le, "Machine learning based bankruptcy prediction of Vietnam companies," *Journal of Economics & Development*, no. 310, pp. 44–53, Apr. 2023, doi: <https://doi.org/10.33301/JED.VI.1066>
- [7] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of Accounting Research*, vol. 4, Supplement, pp. 71–111, 1966, doi: <https://doi.org/10.2307/2490171>
- [8] Y. Wang and M. Campbell, "Financial ratios and the prediction of bankruptcy: The Ohlson model applied to Chinese publicly traded companies," *Proceedings of ASBBS*, vol. 17, no. 1, pp. 334–338, Feb. 2010. [Online]. Available: https://www.researchgate.net/publication/228429542_FINANCIAL_RATIOS_AND_THE_PREDICTION_OF_BANKRUPTCY_THE_OHLSON_MODEL_APPLIED_TO_CHINESE_PUBLICLY_TRADED_COMPANIES
- [9] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *European Journal of Operational Research*, vol. 274, no. 2, pp. 743–758, 2019, doi: <https://doi.org/10.1016/j.ejor.2018.10.024>
- [10] M. H. Tinoco and N. Wilson, "Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables," *International Review of Financial Analysis*, vol. 30, pp. 394–419, 2013, doi: <https://doi.org/10.1016/j.irfa.2013.02.013>