Why Distress Happens: Insights from Explainable Machine Learning in Vietnam

Abstract

This study aims to enrich the field of financial distress prediction for Vietnam enterprises when corporate risk management is highly accentuated. As such, we apply machine learning models, including Single Vector Models, Neural Networks, Random Forest, LightGBM, XGBoost and NGBoost with a sample of 1177 publicly listed firms during the period of 2020-2023 to predict firms' probability of financial distress with 16 chosen financial variables. We then use Shapley Additive Explanations (SHAP) run by XGBoost to explain the specific financial situation of 03 test firms. Our study evidences that the Liquidity and Valuation ratios are significant indicators for the risk of corporate financial distress, which is different from the findings of many studies. We contribute to the current literature in the following aspects. Firstly, we find that two models XGBoost and NGBoost give the best performance and highest stability among all models. Secondly, we notice that financial variables not important globally may be important in a firm's specific case. This study can be a suggestion for individual investors and credit institutions to consider using new financial distress prediction models in an emerging market like Vietnam, where traditional models with huge limitations are still utilized in credit risk management.

1. **INTRODUCTION** (1269 words)

Over the past decades, financial distress prediction has been an important topic of finance that gains significant interest from many researchers. Initially, statistical models were used to predict financial distress based on financial ratios. After that, machine learning techniques have been developed and applied in many studies. However, there is no consistent theory that one specific technique is better than another (Hung and Chen, 2009; Wang *et al.*, 2014). Through empirical studies, researchers across various countries persist in seeking the most effective way to assess the financial health of a firm. Overall, increasing model accuracy, choosing appropriate variables, improving the prediction time and interpreting models continue to be main problems in financial distress prediction. Bellovary *et al.* (2007) reviewed 165 financial distress studies and an interesting question has been raised 'why do we continue develop new techniques in financial distress prediction while many of which achieved good performance?'. The researchers suggested that future research should investigate in the application of these models and refine them if necessary.

The most challenges aspect of machine learning models are their lack of interpretability. In Viet Nam, banks and credit institutions have to filter the credit requirers and clarify their assessment of financial risks for corporate borrowers. However, du Jardin (2018) concluded that artificial intelligent cannot be used as a tool for decision-making because they cannot be explained. Besides, Son *et al.* (2019) implied that while machine learning

models' accuracy is not significantly higher than statistical models, their results cannot be interpreted. This leads to the reason why they are not used in practice. The most challenging aspect of machine learning models is their lack of interpretability. In Viet Nam, banks and credit institutions have to filter the credit requirers and clarify their assessment of financial risks for corporate borrowers. However, du Jardin (2018) concluded that artificial intelligent cannot be used as a tool for decision-making because they cannot be explained. Besides, Son *et al.* (2019) implied that while machine learning models' accuracy is not significantly higher than statistical models, their results cannot be interpreted. This leads to the reason why they are not used in practice.

As a result, many studies have been conducted to explain machine learning models using interpretability methods. Overall, there are three kinds of explainable method: interpretable models, neural network interpretation and model-agnostic methods. In recent years, many studies have used model-agnostic methods such as Local Interpretable Model-agnostic Explanations (LIME), Partial Dependence Plot (PDP), and Shapley Additive Explanations (SHAP). Among these techniques, SHAP stands out due to its strong theoretical foundation and ability to explain clearly the machine learning models at both local and global levels. Currently, there are studies of machine learning corporate financial distress prediction models using SHAP (Perboli and Arabnezhad, 2021; Sigrist and Hirnschall, 2019; Zhang *et al.*, 2022). However, these studies remain some gaps. Some studies compare the efficiency of different models and use SHAP to explain the importance of features to the prediction result but the local explanation for each ratio in a company has not been examined. Moreover, financial ratios in some studies were not clearly identify, leading to no further comparison with financial distress theories.

In Vietnam's emerging economic landscape, enterprises are confronted with huge risk of financial challenges. Vietnam is a rapidly developing economy, which faces challenges such as bad debts and market volatility. Besides, Vietnam also has some specific factors such as dependence on exports and foreign investment, along with changes in financial policies. These characteristic makes forecasting financial distress in Vietnam different from others. According to General Statistics Office of Vietnam, in 2023, there are 89,060 businesses that temporarily ceased operations due to financial difficulties, increases 20.7% compared to 2022. As figure 1, the number of financial distress firms in Vietnam increased significantly in the past 8 years and these financial distress situations underscore the critical need for effective solutions.

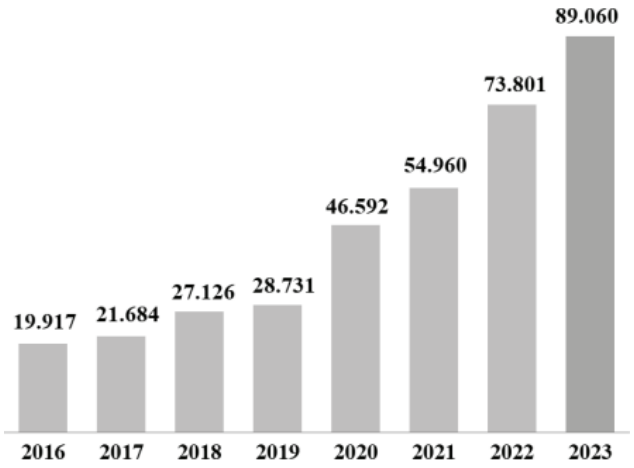


Figure 1: Number of businesses temporarily ceased operations due to financial distress between 2016 and 2023, according to General Statistics Office of Vietnam

To address knowledge gap, our study will explain both the individual impacts and the combined impacts of the financial ratios on the predictive ability of the model and compare the results with the previous studies. Secondly, we assess the performance of machine learning compared to traditional models, giving the reasons to believe in applying machine learning models of financial distress prediction in Vietnam. Finally, we clarify our variables, explain the performance of these factors using financial theories and compare with previous studies.

Overall, this paper used both machine learning and traditional models for financial distress prediction with a sample of 1177 listed firms from 6 sectors (Industry, Mining, Logistic, Forestry and fisheries, Trade and services, Construction) over the period of 2020-2022, with a total of 3531 firms to predict the financial situation of the firm in 2023.

Our study has significant contributions to both theorical foundation and Vietnam economy. Firstly, we analyze financial distress prediction based on the overall importance of ratios with SHAP visualization. Secondly, our study using SHAP to investigate in impacts of financial ratios in each firm. Finally, by clarifying important factors, our study both improves transparency of predicting models and lays the foundation for intergrating modern techniques into financial risk management system.

The next part of this paper is structured as follow: Section 2 gives an overview of previous studies relating to financial distress prediction. Section 3 gives a detailed explanation of the data, while section 4 explains about the methodology of our research. Section 5 shows our result discussion, and Section 6 gives the final conclusion for our study.

2. LITERATURE REVIEW

Since the beginning of research on financial distress prediction, improving model accuracy has always been the main objective of many researchers. The first statistical finding in financial default prediction was the proposal of (Beaver, 1966) which suggested that using multiple ratios offered a better prediction than using only one ratio. Two years later, (Altman, 1968) heeded Beaver's proposal and first utilized multiple discriminant analysis (MDA) models into bankruptcy prediction with higher accuracy. However, they are restricted by statistical assumptions and to overcome this shortcoming, Ohlson (1980) pioneered a more interpretable and superior accurate method, Logistic Regression (LR), to avoid multicollinearity among explanatory variables. The next milestone of financial distress prediction was the emergence of single machine learning models to improve model accuracy, with neural networks (NN) as the first one (Bellovary *et al.*, 2007). Following this, Support Vector Machines (SVM) and Decision Trees (DT) were two other single models applied in financial distress prediction. As developers tried to improve the model's performance, they decided to ensemble the models – the basis for ensemble models, and Random Forest model was first found in 2001. The next decade witnessed the appearance of research-favorite machine learning models, Gradient Boosting, with XGBoost (2014), LGBM (2017) and NGBoost (2019) with its support for intepretability.

However, researchers are still having contradictory opinions about the model efficiency. Wang *et al.* (2014) finds no basis to conclude which model is best for financial distress prediction in every situation. Altman *et al.* (2020) found that LR and NN give better results than SVM, gradient boosting and decision trees. Alfaro *et al.* (2008) showed that over the past decade, ensemble methods have yielded better results than single models. Meanwhile, Jones *et al.* (2017) advocated the adoption of ensemble methods for their significantly better performance than any other classifier in both the cross-sectional and the longitudinal test samples.

We present a compilation of the latest machine learning-based studies in Appendix A. The number of models used in these studies is from 1 to 5 and the most used ensemble and single models are respectively XGBoost and Neutral Networks. The number of independent variables used in those studies is usually from 1 to 37; some researchers use fewer than 20 independent variables but still bring a high classification accuracy (Sigrist and Hirnschall, 2019). Although selecting financial variables is important, there has not been a consensus for the best selection method until now (Balcaen and Ooghe, 2006). In Table 1, the most common variable types are accounting variables, market-based variables and financial ratios. Those variables are mainly divided into 6 main categories: efficiency, liquidity, leverage, profitability, cashflow and valuation. Appendix B gives a detailed description of the variables used in previous research. As predicting the risk of

financial distress of the company is deeply grounded in evaluating a firm's ability to meet its obligations (Zavgren, 1985), variables reflecting efficiency and liquidity, such as Current Ratio (CR) and Asset Turnover Ratio (ASSETURN) are commonly used in research.

Another objective in financial distress study is how to interpret prediction using machine learning models; hence, many scholars have explored various methods to improve the interpretability of these models such as Local Interpretable Model-agnostic Explanation (LIME), Quantitative Input Influence (QII), SHAP and so on. Among these, SHAP stands out due to its strong theoretical foundation and ability to explain clearly the machine learning models at both local and global levels. SHAP has been applied in some environmental domains such as drought prediction (Dikshit and Pradhan 2021) and air quality prediction (Vega Garcia and Aznarte 2020). However, studies using SHAP in financial problems remain rare, specifically, financial distress prediction has not been examined clearly. Perboli and Arabnezhad (2021) have explained how machine learning models predicted a local firm to be financial distress. However, their financial ratios were not clearly identified, leading to no further comparison with financial distress theories. Tran *et al.* (2018) used SHAP to explain financial distress in Viet Nam, however, their studies did not describe how a variable affected the prediction. In the same way, Okay *et al.* (2021) provide SHAP illustration for variables contributed most for the prediction but did not explain how a particular factor affected the results.

Based on the above literature review, our study focuses on Vietnam, an example for a transition economy, to predict the corporate financial distress due to the following reasons. Firstly, although financial distress prediction studies using machine learning methods have attracted increasing attention in many transition economies in decades (Chen *et al.*, 2006; Sun *et al.*, 2011; Wang and Li, 2007; Zheng and Yanhui, 2007), there are relatively few studies in Vietnam. Secondly, interpreting financial distress probability through machine learning models based on SHAP has not been popular in Vietnam and these existing studies have not yet explained the relationship between financial ratios and predicted outcomes clearly. Additionally, Vietnam is considered as a dynamic market when having experienced some economic distress periods such as financial crisis in 2008, macroeconomic fluctuation in 2012-2014 period and COVID 19 pandemic. Therefore, predicting financial distress in listed companies in Vietnam, an emerging market economy, with the application of SHAP method is essential in making investment decisions and lending decisions of other countries in the world.

Compared to prior work, our main contributions to provide a comprehensive analysis for financial default prediction are as follows.

- Our study will predict financial distress based on the overall importance of variables, clarify the relationship between each variable and financial difficulty

probability using financial theories, visualization tools of SHAP, then compare them with findings of previous scholars.

- We use NGBoost to predict financial distress and XGBoost to analyze financial ratios in each firm in our sample data.
- Our paper will compare the performance of machine learning and traditional models, giving the reasons to believe in applying ensemble models of financial distress prediction in Vietnam.

3. **DATA AND VARIABLES** (1904 words)

3.1. Sampling methods and variables

Our study examines a dataset of 1177 publicly listed Vietnamese companies over the period from 2020 to 2023 which still remain actively traded on one exchange HNX, HOSE or UPCOM. Those were not the super-small-sized companies, following the method of Altman (1968). We captured a diverse range of non-financial manufacturing sectors enlisted in Table 3.1.

Table 1: Summary of statistics data by industry

Sector	**Industry**	**Mining**	**Logistic**	**Forestry and fisheries**	**Trade and services**	**Construction**
Number of companies	416	77	109	79	291	205
Percentage	35,34%	6,54%	9,26%	6,71%	24,72%	17,42%

Following Nguyen *et al*.(2023), our data is divided into a training set and a test set, with a ratio of 80-20. With our dataset, 943 observations are set for training and 236 remaining for testing.

We use 16 independent variables as showed in Table 3.2 below, with 3 first variables from the accounting records and 13 following financial ratios. We also have Appendix C show the descriptive statistics for all independent variables from 1-3 years before financial distress. In general, non-financial distress businesses always have a higher median profitability than failed firms (RETA, EBITTA). One year before financial distress, the symptom becomes clearer. Moreover, while financial distress firms have low average liquidity ratios (CR), non-financial difficulty firms show their better performance by higher mean liquidity. Besides, this study tested the correlation between independent variable and get the outcomes as Appendix D. In general, almost autocorrelation values

are lower than 0.8, only correlation between CR and WCTA is higher than 0.8 (0.82) in 3 year pre-distress. These indicates that, data used in this study is suitable for SHAP explanations.

Table 2: Independent Variables

No.	Name	Variable Name	Type
1	Retained Earnings/Total Assets	RETA	Profitability
2	Earning before Interest, Taxes/ Total Assets	EBITTA	Profitability
3	Book value of Equity/ Total Liabilities	BVETL	Valuation
4	Asset Turnover Ratio	ASSETURN	Efficiency
5	Current Ratio	CR	Liquidity
6	Short-term Debt to Asset	STDTA	Leverage
7	Long-term Debt to Asset	LTDTA	Leverage
8	Operating Cash Flow/ Capital Expenditure	OCFCAPEX	Cashflow
9	Enterprise Value/ Revenue	EVR	Valuation
10	Market Value of Equity/ Total Liabilities	MVETL	Valuation
11	Working Capital/ Total Assets	WCTA	Liquidity
12	The Account Payable for Supplier	SUPPAY	Efficiency
13	Financial Leverage	FL	Leverage
14	Price to Earning	BPE	Valuation
15	Operating Cash Flow/ Total Liabilities	OCFTL	Cashflow
16	EBIT/ Interest Coverage	EBITIC	Liquidity

3.2. Firm classification method (199 words)

We follow the classification with Emerging Market Score (EMS), the Z''-score version (Altman, 2005) applied to rank the Mexican corporate bonds in this market:

$$EMS\ Score = 6.56\,X_1 + 3.26\,X_2 + 6.72\,X_3 + 6.56\,X_4 + 3.25$$

where X_1= Working Capital/ Total Assets; X_2=Retained Earnings/ Total Assets
X_3= Operating Income/ Total Assets; X_4= Book Value of Equity/ Total Liabilities

The EMS Score standard for the possibility of financial distress:

- EMS Score $>$ 5.85: Safe zone, in which firms have healthy finances or no risk of bankruptcy. Firms will be classified as "Non-Distressed".
- EMS Score $\leq$ 5.85: Both grey zone and danger zone, in which firms stand from low to high risk of a financial distress. Firms will be classified as "Distressed"(Harrison, 2005).

We use the EMS Score in 2023 to classify the companies into distressed and non-distressed ones, then use the 1 year, 2 years and 3 years pre-distress model (with data from 2020-2022 and classification in 2023) to make forecast for the financial situation of the company.

4. METHODOLOGY (660 words)

4.1. Modeling Method

In this study, we use 8 models of three groups: traditional models with Discriminant Analysis (Altman, 1968) and Logistic Regression (Ohlson, 1980), single machine learning models with Support Vector Machine (Cortes and Vapnik, 1995) and Neural Network (Odom and Sharda, 1990), and ensemble machine learning models with Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), LightGBM (Ke *et al.*, 2017) and NGBoost (Duan *et al.*, 2020).

4.1.1. Discriminant Analysis

Discriminant Analysis was used by Altman (1968) to clasify firms into 2 groups using the Z-Score function as follows:

$$Z = w_0 + \sum_{i=1}^{n} w_i x_i$$

where w_i are calculated by minimizing within-class variances and maximizing between-class variances, x_i are the independent variables.

4.1.2. Logistic Regression

Logistic regression is a binary classification method, which uses the sigmoid function to transform a linear independent variable into a value between 0 and 1. The sigmoid function is defined as follows:

$$Z = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{p}\beta_i x_i)}}$$

where x_i are the independent variables, β_0 is the intercept and and β_i are the regression coefficient.

4.1.3. Support Vector Machine (SVM)

Support Vector Machine is one of the first machine learning algorithm for binary classification. It aims to find the best hyperplane in space of many features. This hyperplane aims to maximize the Function Margin, which is the distance between hyperplane and the data points of each class. Function Margin can be defined as follows:

$$FM = y(wx + b)$$

where y is class label, w is the weight vector, x is the vector of input data, b is the bias term.

4.1.4. Neural Network

Neural networks is a machine learning algorithm which mimic human decision-making process by using interconnected neurons and activation functions to train input data, capture complex relationships and make predictions. The output y_j^l of a neuron j in layer l is defined as:

$$y_j^l = \sigma\left(\sum_i x_{ij}^{(l)} o_i^{(l-1)} + b_j^{(l)}\right)$$

where $x_{ij}^{(l)}$ is the weight connecting neuron i in layer $l-1$ to neuron j, $b_j^{(l)}$ is the bias of neuron j in layer l , $o_i^{(l-1)}$ is the output of neuron i in previous layer.

4.1.5 Random Forest

Random Forest is an algorithm to combine multiple decision trees with randomized data sample and feature selection to improve prediction accuracy and reduce overfitting. The class assigned by the decision tree can be calculated as follows:

$$C_i = \arg arg\,max_c \sum_{i=1}^{N} I(C_i = c)$$

where C_i is the class by the $i - th$ decision tree, $I(C_i = c)$ is the sum of occurences across all decision trees.

4.1.6. XGBoost

XGBoost, or eXtreme Gradient Boosting, is a prominent ensemble tree model which based on gradient boosting principles to enhance predictive accuracy through an ensemble of weak learners, primarily decision trees, while addressing challenges related to bias and variance.

The main predictive model of XGBoost can be defined as follows:

$$F(x) = \sum_{m=1}^{M} f_m(x)$$

where $F(x)$ is the financial prediction for the input x, and $f_m(x)$ represents the output function of each individual tree in the ensemble. XGBoost is of higher speed and efficiency than traditional Gradient Boosting due to its regularized learning objective to prevent overfitting, which can be denoted in the loss function as follows:

$$L(F) = \sum_{i=1}^{N} l\left(y_i, F\left(x_i\right)\right) + \sum_{m=1}^{M} W(f_m)$$

with the regularization function:

$$W(f_m) = \gamma M + \frac{1}{2} l \sum_{m=1}^{M} \|w_m\|^2$$

Where l is the loss function that mesures the difference between the predicted and real values, W is the regularization function calculated based on parameters γ, l, the number of leaves M, and the score w at each leaf.

4.1.7. LightGBM

LightGBM is an advanced gradient boosting decision tree algorithm, which incorporates Gradient-Based One-Side Sampling (GOSS) for optimal node and Exclusive Feaute Bundling (EFB) to handle dense features effectively. This model helps enable vertical tree growth, improving training speed and accuracy, esspecially for large scale datasets.

4.1.8. NGBoost

NGBoost is a newly designed probabilistic forecasting algorithm that is a variant of the gradient boosting framework. In contrast to other algorithms, it does not provide the

expected value in the case of regression but rather the conditional probability of the output. When it comes to classification problems, NGBoost like most classification algorithms returns the probability of any given class.

4.2. Model and tunning parameters choosing

To get the best performance of each model, we need to select the optimal combination of model hyperparameters. In Cross-Validation step, we use Gridsearch to choose trained parameters with a 10-fold cross-validation, then rerun the set with another 10-fold cross-validation in a test sample. We use the accuracy rate to assess the trained model performance, and this process will stop when every dataset is tested. All the optimal hyperparameters for lgbm, XGBoost, NGBoost, SVM, RF, Neural Networks are put at Appendix E.

We evaluate the performance of the final models by the prediction accuracy, Type-I error, Type-II error, receiver operation characteristics (ROC) and area under the ROC curve (AUC), in which:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Companies\ in\ the\ Test\ set}$$

Type-I Error (or False Positive) is the misclassification of non-distressed companies as the distressed companies, while Type-II Error (False Negative) is the misclassification of distressed firms as the normal ones. AUC-ROC measures the performance of the classification models. Graphically, ROC stays in a coordinate system with y-axis and x-axis respectively as True Positive Rate and False Positive Rate. AUC is the area under the ROC curve showing how well the model is classifying. AUC ranges between 0 and 1, and the acceptable level of AUC is between 0.7 and 0.8, 0.8 to 0.9 is excellent, and 0.9+ is outstanding (Mandrekar, 2010).

4.3. SHapley Additive Explanations

As machine learning models evolve at the expense of understandability, Shapley Additive Explanations (SHAP) is used to explain the complexities of the decision-making processes behind the algorithms. First introduced by Lundberg and Lee (2017), SHAP is a model-independent outstanding method for a strong mathematical foundation and proven consistency. Therefore, this study uses SHAP to interpret the complex decision-making processes of ensemble methods.

SHAP attaches Shapley values to all features to quantitatively measure their importance on the payout, thus point out the most important features. SHAP also shows the additive decomposition – how the contribution of all features is linearly added to generate the prediction result. The Shapley values are represented as an additive feature attribution method and can be expressed as follows:

$$g(x') = \alpha_0 + \sum_i^M \alpha_i x_i'$$

$g(x')$ is the explanation model, x' is the coalition vector $(0;1)^M$, α_0 is the average prediction of the model over all feature subsets; M is the number of input variables; α_i is Shapley value for the i_{th} feature.

The formula of Shapley value f_i for each feature:

$$f_i := f_i(f,x) = \sum_{S \acute{I} F \backslash \{i\}} \frac{|S|!(F-|S|-1)}{F!} [f_x(S \grave{E} \{i\}) - f_x(S)]$$

where F is the set of all features, $|S|$ represents the number of variables in variable subset S when excluding x_i. For each subset S, we see the difference between two predictions: $f_x(S \grave{E} \{i\})$ and $f_x(S)$. $f_x(S \grave{E} \{i\})$ is the model's prediction when x_i is included in subset S while $f_x(S)$ represents the model's prediction features in subset S. From the formula, there are $|S|!\,(F - |S| - 1)$ possible orderings of the features i and F! possible orderings of all features in set F.

To evaluate the overall important, the average of the absolute Shapley values across all observations is defined as follows:

$$I_i = \frac{1}{n} \sum_{j=1}^{n} \left| f_i^{(j)} \right|$$

Where I_i is the feature importance for the i-th variable, n is the total number of observations, and $f_i(j)$ is the Shapley value of the variable x_i in the j-th observation.

In SHAP explanation, we access the relationship between the explanatory and target variable throughout all the features with SHAP Summary Plot. Moreover, to explain the local effects of a feature on the prediction, we use SHAP Dependence Plot. Finally, SHAP force plot is applied to explain why a particular company is predicted to be financial distress by machine learning model. Through SHAP, we can explain the non-monotonic effects of a feature on the model to decide whether a relationship is non-linear. Another reason to trust SHAP is its inclination to show the extent of linkage between the features and the outcome without modelling the causality. This is because Machine Learning algorithms only learn to seek out the associations through data mining, and changing the data will change the patterns of associations.

5. RESULTS

5.1. Performance assessment of the models

	Ensemble models				Single Models		Trad. models	
	LGBM	**XGBoost**	**NGBoost**	**RF**	**SVM**	**NN**	**DA**	**LR**
1Y	90%	88%	89%	88%	89%	88%	89%	83%
2Y	86%	84%	85%	86%	86%	82%	82%	81%
3Y	83%	83%	85%	81%	79%	82%	80%	75%

Table 3: Correct Classification of Models

The classification accuracies are listed in Table 5.1. Overall, all models perform really well in the short run (except for LR) with an accuracy of 88-90%. However, for longer prediction horizon, ensemble models perform much better than single models and traditional models, except for SVM. This finding accords with what du Jardin (2015) found about the efficiency of statistical models. Among three ensemble models, NGBoost and XGBoost are the most two stable models in terms of classification accuracy.

		Ensemble models				Single Models		Trad. models	
	Number of errors	**LGBM**	**XGBoost**	**NGBoost**	**RF**	**SVM**	**NN**	**DA**	**LR**
1Y	Type-I Error	12	12	8	10	11	17	15	26
	Type-II Error	20	17	17	18	15	12	12	15

2Y	Type-I Error	19	13	14	12	13	15	19	12
	Type-II Error	14	25	22	22	25	27	24	32
3Y	Type-I Error	22	16	14	19	20	18	20	26
	Type-II Error	18	25	22	26	30	24	28	32
Average	Type-I Error	17,6	13,6	12	13, 6	14,6	16, 6	18	21, 3
	Type-II Error	17,3	22,3	20,3	22	23,3	21	21, 3	26, 3

Table 4: Type-I Error and Type-II Error

Table 5.2 shows the number of Type-I Errors and Type-II Errors in our test set. The disparities between the two errors follow the results of Alfaro *et al.* (2008). Almost all models make fewer Type-I Errors than Type-II Errors, which means that the prediction for firms at risk of financial distress is better than the prediction for financially decent firms. In the case of financial distress, making a Type-II Error causes more serious damage to the financial system than making a Type-I Error (Altman *et al.*, 1977). Ensemble models are found to stably make fewer Type-II Errors than both single models and traditional models along the observing period. Therefore, in terms of both Type-I and Type-II Errors, we can conclude that ensemble models are the most advisable to best avoid Type-II Errors.

	Ensemble models				**Single Models**		**Trad. models**	
AUC	**LGBM**	**XGBoost**	**NGBoost**	**RF**	**SVM**	**NN**	**DA**	**LR**
1Y	96%	96%	96%	96%	96%	96%	96%	95%
2Y	92%	93%	92%	92%	90%	89%	91%	87%
3Y	90%	90%	92%	90%	88%	87%	89%	85%

Table 5: AUC (Area under the ROC curve) of models

The AUC of 8 models is shown in Table 5.3. The financial distress prediction of every model is most efficient for one year pre-distress, at about 95-96%. The ensemble models here are found to outperform traditional models in the longer term and classify slightly better with higher stability than the single models. The findings relating to the efficiency of XGBoost align with the research of (Carmona *et al.*, 2019).

5.2. Visual description of the most important features to the result prediction

We use two powerful tools of SHAP to visualize the relationship between the explanatory variables and the target variable: SHAP Feature Importance, SHAP Summary Plot and SHAP Dependence Plot. With SHAP Feature Importance, features are ranked from most impactful to the least impactful to the target variable shown on the vertical axis, based on the mean SHAP value in every situation shown on the horizontal axis. Compared to SHAP Feature Importance, SHAP Dependence Plot not only gives the rankings based on the mean SHAP value but also shows the relationship between the magnitude of the ratios to the prediction result. The value of the feature is described by colors with low (blue), middle (purple) and high (red). Lastly, we use SHAP Dependence Plot, which describes the impact of each data point on the prediction result with the specific range of the input data.

According to the SHAP Feature Importance (on the left of Figure 5.1), we suggest that the four respectively ranked most important ratios from 1 year pre-distress model include Current Ratio (CR), Working Capital over Total Assets (WCTA), Book Value of Equity to Total Liabilities (BVETL), Market Value of Equity over Total Liabilities (MVETL). An interesting finding here is the important ratios for financial distress prediction are relatively consistent for different prediction periods, namely Current Ratio, Working Capital/ Total Assets, Book Value of Equity over Total Liabilities and Market Value of Equity over Total Liabilities.

Through the SHAP Summary Plot (on the right side of Figure 5.1), we can both determine the key features for predicting financial distress and see the relationship between feature values and their impacts on the prediction. As expected, from this above plot, current ratio (CR) or working capital over total assets (WCTA) are the two most important features in predicting financial distress. This is because lower CR or WCTA means that a firm has lower liquidity with fewer current assets, hence lower ability to pay off short-term debts. The difficulty in debt payments, therefore, can raise the risk that a firm faces financial distress in their operation. This result is supported by several previous research such as (Beaver, 1966; Lokanan and Sharma, 2024; Pham Vo Ninh *et al.*, 2018).

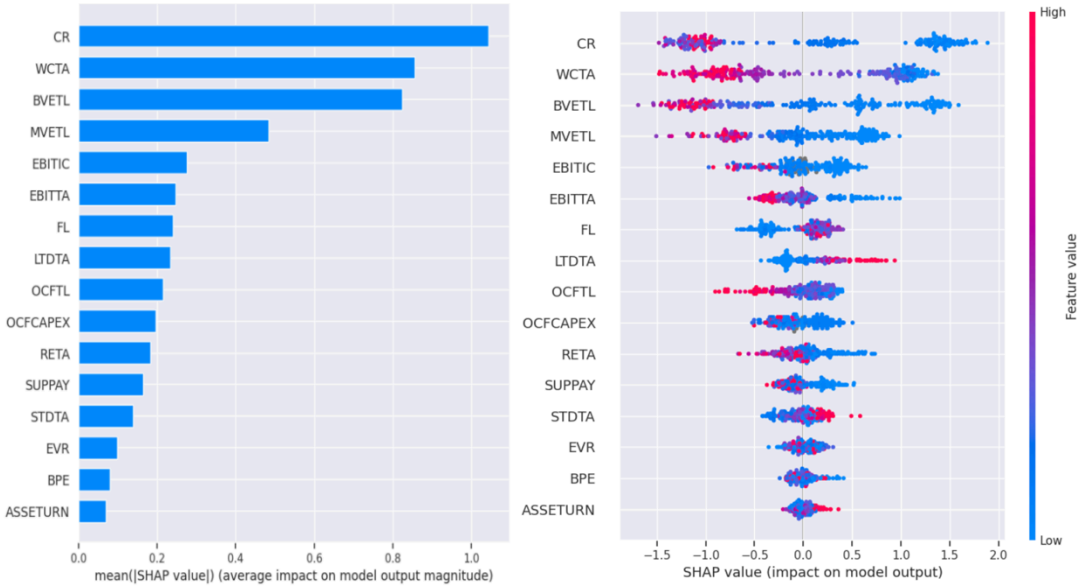


Figure 5.1: SHAP Feature Importance and SHAP Summary Plot

We also analyze the impacts of two ratios (CR and WCTA) on financial distress in more detail through the SHAP Dependence Plot (Figure 5.2). With CR ranging from 1 to 5, the threshold of current ratio in our dataset that a company facing higher possibility of financial distress is about 1.5. The WCTA of firms here usually ranges from -0.1 to 0.5, and if the WCTA exceeds 0.15, the firm then undergoes lower risk of financial difficulties.

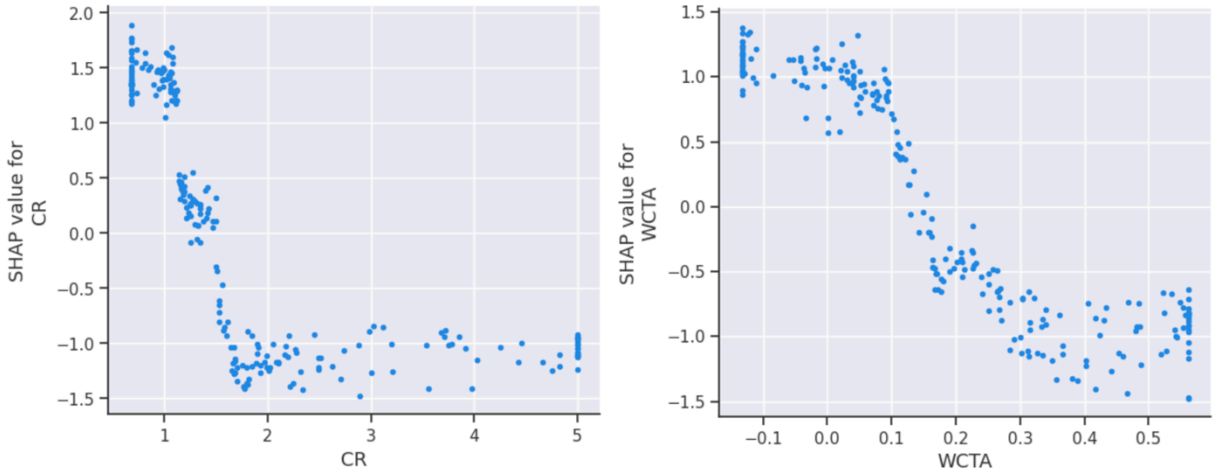


Figure 5.2: SHAP Dependence Plot of CR, WCTA in XGBoost 1 year pre-distress

The third and fourth important ratios are Book Value of Equity over Total Liabilities (MVETL) and Book Value of Equity over Total Liabilities (MVETL). Our analysis of the SHAP summary plot suggests that a high book value of equity over total liabilities (BVETL) (depicted on the

right side of Figure 5.3) corresponds to a lower Shapley value, or a lower probability of facing financial difficulties. This result was supported by Altman (1968) and Fama and French (1992). It is explained that a higher BVETL ratio represents a firm's better financial capability of paying debts and less dependence on debt in a firm's capital structure, which indicates the decent financial situation of a company. The similar relationship is also seen in the market value of equity over total liabilities (MVETL) (on the right side of Figure 5.3) - another index for valuation. Altman (1968) showed that a firm with a high MVETL would also have higher tolerance against financial distress for two reasons. At first, when a firm is highly valuated, it will have higher capability to finance its debts with its own equity. Besides, a firm with a high MVETL would also have higher creditworthiness to investors and credit suppliers, who can give the firms external financial resources with favorable conditions to finance its debt. Both those two reasons explain why a firm with a higher MVETL faces a smaller risk of financial distress.

We also find out the relationship between the two above ratios (BVETL and MVETL) and the financial distress probability through the SHAP Dependence Plot (Figure 5.3). In our dataset, the risk of financial distress for the firm reduces significantly when the Market Value of Equity and Book Value of Equity are valuated to be equal or higher than its debts. In other words, the threshold of BVETL and MVETL for a firm to start facing a financial distress is around 1. SHAP Dependence Plots for other ratios are put at Appendix F.

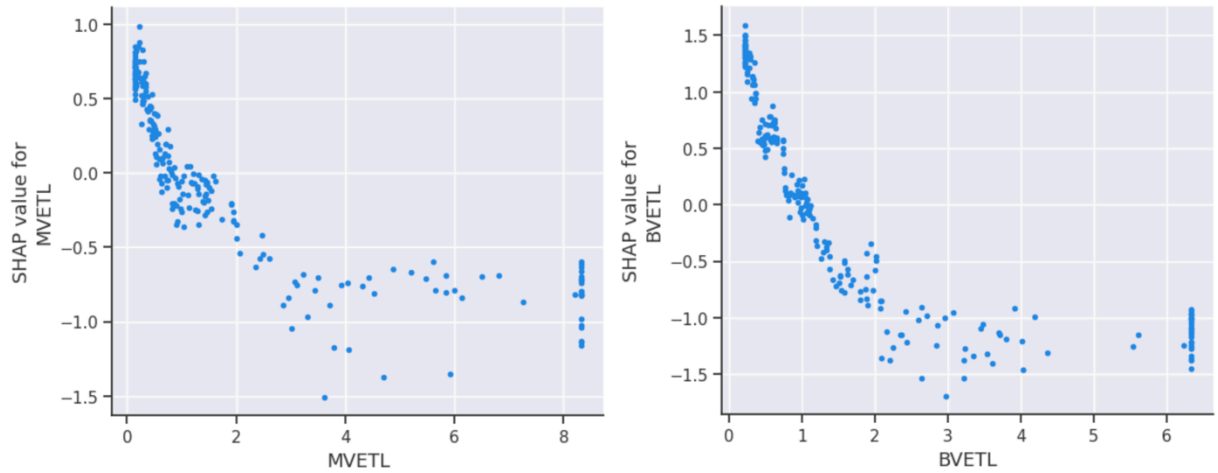


Figure 5.3: SHAP Dependence Plot of MVETL, BVETL in XGBoost 1 year pre-distress

5.3. What puts a company at financial risk? SHAP Interpretations of Financial Distress Models

The most important feature of SHAP is to explain why XGBoost predicts that a company faces financial distress when allowing us to explain the financial distress prediction for every company in the dataset. Each feature is illustrated as an arrow, with color scheme used to indicate the contribution to the financial distress predictions. The red color indicates a contribution to increase the risk of financial distress, while blue

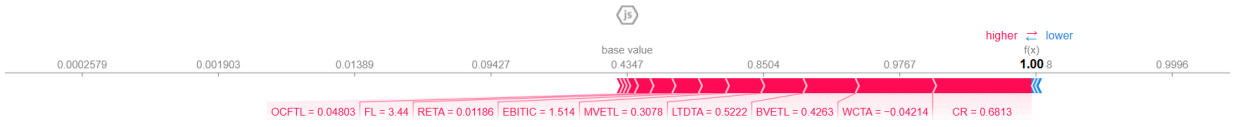
color signifies an impact to reduce the financial distress. Features with a more significant contribution are closer to the dividing boundary between red and blue. The size of a bar represents the extent of impact. Finally, the predicted financial distress probability is calculated by the addition of all values and highlighted in black.

Thanks to SHAP force plot, valuable insights about the relative significance of different factors in predicting a company's likelihood of financial distress can be obtained. This helps to understand how each factor devotes to the overall prediction.

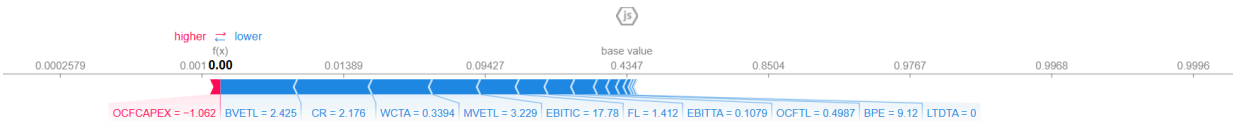
In figure 5.1, CII is predicted to 100% face financial distress. The most crucial features are as follows:

CR (Current Ratio): 0.6813
WCTA (Working Capital/ Total Assets): -0.04214
BVETL (Book Value of Equity/ Total Liabilities): 0.4263
LTDTA (Long-term Debt to Asset): 0.52222
MVETL (Market Value of Equity/ Total Liabilities): 0.3078
EBITIC (Earning before Interest, Taxes/ Interest Coverage): 1.514
RETA (Retained Earning/ Total Assets): 0.01186
FL (Financial Leverage): 3.44
OCFTL (Operating Cash Flow/ Total Liabilities): 0.04803

With SHAP force plot, we can explain why XGBoost predicts that the firm will be financial distress. This firm suffers low liquidity (CR, WCTA) and low profitability (RETA) . Compared to median of 0.12 for distress and 0.03 for non-distress company in Appendix C*6*, this company has a high long-term debt to assets (LTDTA) of 0.52222. In reality, projects primarily financed by liabilities create significant repayment pressure, which can lead to financial dificulties for the business.



Financially-distressed Firm (CII)

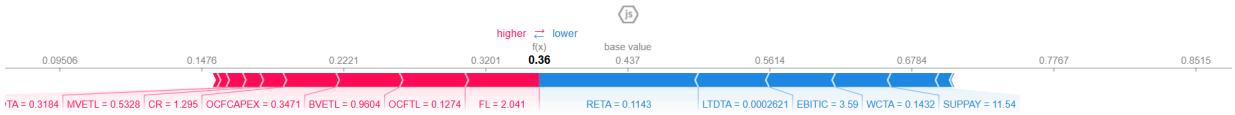


Financially non-distressed Firm (FMC)

In the bottom of the figure, FMC is forecast to avoid financial distress. The significant variables that have strong impacts are presented as follows:

BVETL (Book Value of Equity/ Total Liabilities): 2.425
CR (Current Ratio): 2.176
WCTA (Working Capital/ Total Assets): 0.3394
MVETL (Market Value of Equity/ Total Liabilities): 3.229
EBITIC (Earning before Interest, Taxes/ Interest Coverage): 17.78
FL (Financial Leverage): 1.412
EBITTA (Earning before Interest, Taxes/ Total Assets): 0.1079

BVETL and MVETL have high SHAP values, indicating strong relationship of equity and liabilities. In theory, high MVETL and BVETL ratio suggest that a firm has solid financial foundation, which reduces the risk of financial difficulty. This is proved by the higher average values for non-distress company with MVETL at 9.75 and BVETL at 7.63 in Appendix C. Other factors, such as CR (2.175) and WCTA (0.3394) also has a positive impact, reflecting the high liquidity of this company. According to theory, a high liquidity shows the firm's ability to meet short-term obligations, which reduce the risk of financial distress. In this case, non-distressed companies has a higher average current ratio of 4.41 than the distress group's 1.05 (Appendix C)
Another case of financial distress firm is illustrated in Figure below



Financially Non-distressed Firm (PAN)

The probability of financial distress is 36% with significant features as belows:
RETA (Retained Earning/ Total Assets): 0.1143
LTDTA (Long-term Debt to Asset): 0.0002621
EBITIC (Earning before Interest, Taxes/ Interest Coverage): 3.59
WCTA (Working Capital/ Total Assets): 0.1432
SUPPAY (The Account Payable for Supplier): 11.54

In fact, the 2020-2022 period saw a downward trend in the CII's earnings due to the shock in real estate demand as a result of COVID-19. PAN and FMC also suffered from supply chain disruption and geopolitical instability that made them struggle a bit in 2020. However, the two firms recovered in the next two years due to their penetration into new foreign markets.

The use of financial ratios in financial distress prediction varies across different firms, as in Figure 5.4 and 5.5, the important set of non financial distress company is BVETL, CR, WCTA, MVETL, while RETA, LTDTA, EBITIC are found to significantly contribute to

financial distress of company in Figure 5.6. In addition, SHAP reveals that some factors may not be the most crucial factor globally, but may be significant locally. Because each firm has its own characteristics, the global explanation may not always be compatible with local explanation. Therefore, the SHAP force plot substantially clarifies the benefits of SHAP. Moreover, this adaption can enable practitioners to understand why a firm is financially distressed by identifying the significant features to the prediction. SHAP force plot provides a clear and data-driven way to interpret the outcome of machine learning models in a local context, which helps to take accurate actions to reduce risk of financial distress for each company.

6. Conclusions

In our research, we compared the performance of machine learning models and traditional models on Vietnamese enterprises data from 2020 to 2023. Overall, ensemble machine learning models have the highest accuracy and stability over both short term and long term, and this finding supports previous research that ensemble methods perform better than single and traditional ones. In addition, XGBoost is the most stable among our ensemble models, so we employed SHAP to illustrate the key features and their relation to the prediction of corporate difficulty. Moreover, we analyzed and pointed out the most important ratios along with financial distress probability to each company based on Shapley values assigned to each ratio. This is a significant strength of SHAP method because interpreting the machine learning models in local context could mitigate the risk of financial distress for each firm. Specifically, from the SHAP Global explanation, we found that Liquidity and Valuation Ratios have significant impacts on the financial distress prediction; while regarding SHAP Local explanation, we can find out more that RETA may be vital to predict the risk of financial distress for an individual company.

Our results can bring several practical implications for legal entities. Initially, based on an interpretation method consisting of local and global explanations, credit suppliers, investors, firms, and banking institutions can trust the “black-box” models and better make decisions to give right loans and credit ratings to firms. Specifically, the SHAP Local Explanation gives a particular prediction for each firm; hence, practitioners could be more informed and take actions that are appropriate to an individual corporate. Besides, the intepretability also allows the banks to give clear reasons to the rejected borrowers according to the “Right to Explanation”. Intrinsically, the higher interpretability also allows the CFO and firm managers to implement proper policies for a firm, so as to develop sustainably in the future.

In our paper, we only use 16 financial ratios at a specific point in time, so we suggest that further research can extend in two ways: (1) to explore new non-financial factors that may cause the risk of financial distress, (2) to utilize lagged information to have a deeper insight of the financial trends in a firm. Furthermore, future research can

also aim to explain how correlation among selected variables can be best modelled after the calibration method. Another orientation for future financial distress prediction work is to widen the scope of machine learning models used, like Kernel-SVM, K-Nearest Neighbors (KNN), to CatBoost or AdaBoost, allowing the stakeholders to have more choices for consideration in their risk management process.

Reference list

Alfaro, E., García, N., Gámez, M. and Elizondo, D. (2008), "Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks", *Decision Support Systems*, Vol. 45 No. 1, pp. 110–122, doi: 10.1016/j.dss.2007.12.002.

Altman, E.I. (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *The Journal of Finance*, [American Finance Association, Wiley], Vol. 23 No. 4, pp. 589–609, doi: 10.2307/2978933.

Altman, E.I. (2005), "An emerging market credit scoring system for corporate bonds", *Emerging Markets Review*, Vol. 6 No. 4, pp. 311–323, doi: 10.1016/j.ememar.2005.09.007.

Altman, E.I., Haldeman, R.G. and Narayanan, P. (1977), "ZETATM analysis A new model to identify bankruptcy risk of corporations", *Journal of Banking & Finance*, Vol. 1 No. 1, pp. 29–54, doi: 10.1016/0378-4266(77)90017-6.

Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K. and Suvas, A. (2020), "A Race for Long Horizon Bankruptcy Prediction", *Applied Economics*, Routledge, Vol. 52 No. 37, pp. 4092–4111, doi: 10.1080/00036846.2020.1730762.

Balcaen, S. and Ooghe, H. (2006), "35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems", *The British Accounting Review*, Vol. 38 No. 1, pp. 63–93, doi: 10.1016/j.bar.2005.09.001.

Beaver, W.H. (1966), "Financial Ratios As Predictors of Failure", *Journal of Accounting Research*, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], Vol. 4, pp. 71–111, doi: 10.2307/2490171.

Bellovary, J.L., Giacomino, D.E. and Akers, M.D. (2007), "A Review of Bankruptcy Prediction Studies: 1930 to Present", *Journal of Financial Education*, Financial Education Association, Vol. 33, pp. 1–42.

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45 No. 1, pp. 5–32, doi: 10.1023/A:1010933404324.

Carmona, P., Climent, F. and Momparler, A. (2019), "Predicting failure in the U.S. banking sector: An extreme gradient boosting approach", *International Review of Economics & Finance*, Vol. 61, pp. 304–323, doi: 10.1016/j.iref.2018.03.008.

Chen, J., Marshall, B.R., Zhang, J. and Ganesh, S. (2006), "Financial Distress Prediction in China", *Review of Pacific Basin Financial Markets and Policies*, World Scientific Publishing Co., Vol. 09 No. 02, pp. 317–336, doi: 10.1142/S0219091506000744.

Chen, T. and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, pp. 785–794, doi: 10.1145/2939672.2939785.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks", *Machine Learning*, Vol. 20 No. 3, pp. 273–297, doi: 10.1007/BF00994018.

Duan, T., Anand, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A. and Schuler, A. (2020), "NGBoost: Natural Gradient Boosting for Probabilistic Prediction", *Proceedings*

of the 37th International Conference on Machine Learning, presented at the International Conference on Machine Learning, PMLR, pp. 2690–2700.

Harrison, M.E. (2005), *A Study of Altman's (1983) Revised Four -Variable Z -Score Bankruptcy Prediction Model for Asset Sizes and Manufacturing and Service Companies*, D.B.A., *ProQuest Dissertations and Theses*, Nova Southeastern University, United States -- Florida.

Hung, C. and Chen, J.-H. (2009), "A selective ensemble based on expected probabilities for bankruptcy prediction", *Expert Systems with Applications*, Vol. 36 No. 3, Part 1, pp. 5297–5303, doi: 10.1016/j.eswa.2008.06.068.

du Jardin, P. (2010), "Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy", *Neurocomputing*, Vol. 73 No. 10, pp. 2047–2060, doi: 10.1016/j.neucom.2009.11.034.

du Jardin, P. (2015), "Bankruptcy prediction using terminal failure processes", *European Journal of Operational Research*, Vol. 242 No. 1, pp. 286–303, doi: 10.1016/j.ejor.2014.09.059.

du Jardin, P. (2018), "Failure pattern-based ensembles applied to bankruptcy forecasting", *Decision Support Systems*, Vol. 107, pp. 64–77, doi: 10.1016/j.dss.2018.01.003.

Jones, S., Johnstone, D. and Wilson, R. (2017), "Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks", *Journal of Business Finance & Accounting*, Vol. 44 No. 1–2, pp. 3–34, doi: 10.1111/jbfa.12218.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., *et al.* (2017), "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc.

Lin, K.C. and Dong, X. (2018), "Corporate social responsibility engagement of financially distressed firms and their bankruptcy likelihood", *Advances in Accounting*, Vol. 43, pp. 32–45, doi: 10.1016/j.adiac.2018.08.001.

Lokanan, M. and Sharma, S. (2024), "The use of machine learning algorithms to predict financial statement fraud", *The British Accounting Review*, Vol. 56 No. 6, p. 101441, doi: 10.1016/j.bar.2024.101441.

Lundberg, S. and Lee, S.-I. (2017), "A Unified Approach to Interpreting Model Predictions", arXiv, 25 November, doi: 10.48550/arXiv.1705.07874.

Mandrekar, J.N. (2010), "Receiver Operating Characteristic Curve in Diagnostic Test Assessment", *Journal of Thoracic Oncology*, Elsevier, Vol. 5 No. 9, pp. 1315–1316, doi: 10.1097/JTO.0b013e3181ec173d.

Murugan, M.S. and T, S.K. (2023), "Large-scale data-driven financial risk management & analysis using machine learning strategies", *Measurement: Sensors*, Vol. 27, p. 100756, doi: 10.1016/j.measen.2023.100756.

Nguyen, H.H., Viviani, J.-L. and Ben Jabeur, S. (2023), "Bankruptcy prediction using machine learning and Shapley additive explanations", *Review of Quantitative Finance and Accounting*, doi: 10.1007/s11156-023-01192-x.

Nguyen, M., Nguyen, B. and Liêu, M.-L. (2024a), “Corporate financial distress prediction in a transition economy”, *Journal of Forecasting*, Vol. 43 No. 8, pp. 3128–3160, doi: 10.1002/for.3177.

Nguyen, M., Nguyen, B. and Liêu, M.-L. (2024b), “Corporate financial distress prediction in a transition economy”, *Journal of Forecasting*, Vol. 43 No. 8, pp. 3128–3160, doi: 10.1002/for.3177.

Odom, M.D. and Sharda, R. (1990), “A neural network model for bankruptcy prediction”, *1990 IJCNN International Joint Conference on Neural Networks*, presented at the 1990 IJCNN International Joint Conference on Neural Networks, pp. 163–168 vol.2, doi: 10.1109/IJCNN.1990.137710.

Ohlson, J.A. (1980), “Financial Ratios and the Probabilistic Prediction of Bankruptcy”, *Journal of Accounting Research*, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], Vol. 18 No. 1, pp. 109–131, doi: 10.2307/2490395.

Okay, F.Y., Yıldırım, M. and Özdemir, S. (2021), “Interpretable Machine Learning: A Case Study of Healthcare”, *2021 International Symposium on Networks, Computers and Communications (ISNCC)*, presented at the 2021 International Symposium on Networks, Computers and Communications (ISNCC), pp. 1–6, doi: 10.1109/ISNCC52172.2021.9615727.

Perboli, G. and Arabnezhad, E. (2021), “A Machine Learning-based DSS for mid and long-term company crisis prediction”, *Expert Systems with Applications*, Vol. 174, p. 114758, doi: 10.1016/j.eswa.2021.114758.

Pham Vo Ninh, B., Do Thanh, T. and Vo Hong, D. (2018), “Financial distress and bankruptcy prediction: An appropriate model for listed firms in Vietnam”, *Economic Systems*, Vol. 42 No. 4, pp. 616–624, doi: 10.1016/j.ecosys.2018.05.002.

“Sach-trang-doanh-nghiep-Viet-Nam-2024.pdf”. (n.d.). .

Sigrist, F. and Hirnschall, C. (2019), “Grabit: Gradient tree-boosted Tobit models for default prediction”, *Journal of Banking & Finance*, Vol. 102, pp. 177–192, doi: 10.1016/j.jbankfin.2019.03.004.

Smith, M. and Alvarez, F. (2022), “Predicting Firm-Level Bankruptcy in the Spanish Economy Using Extreme Gradient Boosting”, *Computational Economics*, Vol. 59 No. 1, pp. 263–295, doi: 10.1007/s10614-020-10078-2.

Son, H., Hyun, C., Phan, D. and Hwang, H.J. (2019), “Data analytic approach for bankruptcy prediction”, *Expert Systems with Applications*, Vol. 138, p. 112816, doi: 10.1016/j.eswa.2019.07.033.

Sun, J., Jia, M. and Li, H. (2011), “AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies”, *Expert Systems with Applications*, Vol. 38 No. 8, pp. 9305–9312, doi: 10.1016/j.eswa.2011.01.042.

Tran, K.L., Le, H.A., Nguyen, T.H. and Nguyen, D.T. (2022), "Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam", *Data*, Multidisciplinary Digital Publishing Institute, Vol. 7 No. 11, p. 160, doi: 10.3390/data7110160.

Veganzones, D. and Séverin, E. (2018), "An investigation of bankruptcy prediction in imbalanced datasets", *Decision Support Systems*, Vol. 112, pp. 111–124, doi: 10.1016/j.dss.2018.06.011.

Wang, G., Ma, J. and Yang, S. (2014), "An improved boosting based on feature selection for corporate bankruptcy prediction", *Expert Systems with Applications*, Vol. 41 No. 5, pp. 2353–2361, doi: 10.1016/j.eswa.2013.09.033.

Wang, Z. and Li, H. (2007), "Financial distress prediction of Chinese listed companies: a rough set methodology", *Chinese Management Studies*, Emerald Group Publishing Limited, Vol. 1 No. 2, pp. 93–110, doi: 10.1108/17506140710758008.

Zavgren, C.V. (1985), "Assessing the Vulnerability to Failure of American Industrial Firms: A Logistic Analysis", *Journal of Business Finance & Accounting*, Vol. 12 No. 1, pp. 19–45, doi: 10.1111/j.1468-5957.1985.tb00077.x.

Zhang, Z., Wu, C., Qu, S. and Chen, X. (2022), "An explainable artificial intelligence approach for financial distress prediction", *Information Processing & Management*, Vol. 59 No. 4, p. 102988, doi: 10.1016/j.ipm.2022.102988.

Zheng, Q. and Yanhui, J. (2007), "Financial Distress Prediction Based on Decision Tree Models", *2007 IEEE International Conference on Service Operations and*

Logistics, and Informatics, presented at the 2007 IEEE International Conference on Service Operations and Logistics, and Informatics, pp. 1–6, doi: 10.1109/SOLI.2007.4383925.

Appendix A

Studies	Models	Number of ratios	Sample size	Variable types	Interpretable Method
(du Jardin, 2010)	NN	14	1020	Financial Ratios	None
(du Jardin, 2015)	NN, SOM	36	18620	Financial Ratios and Market-based Variables	None
(Jones *et al.*, 2017)	NN, AdaBoost, SVM, RF	35	30129	Financial ratio	Relative variable importance
(Pham Vo Ninh *et al.*, 2018)	DD	10	6736	Accounting variables, Market – based variables, macroeconomic variables	None
(Veganzones and Séverin, 2018)	NN, RF, SVM	10	1500	Financial ratios	None
(Perboli and Arabnezhad, 2021)	RF, XGBoost, NN	15	8959	Financial ratios	SHAP
(Smith and Alvarez, 2022)	NN, RF, XGBoost, SVM, LightGBM	17	64057	Financial ratios and balance-sheet variables	Feature importance score and Partial dependence plot
(Tran *et al.*, 2022)	SVM, DT, RF, XGBoost, NN	25	3277	Financial ratios and market factors	SHAP
(Lokanan and Sharma, 2024)	RF, NGBoost, CatBoost, DT,	37	3954	Accounting and Financial Ratios	None
(Murugan and T, 2023)	NN, XGBoost	10	Large-scale Data	Accounting and Financial ratios	None
(Nguyen *et al.*, 2024a)	SVMs, NN, RF, DT	11	12685	Financial Ratio, Market-based Variables, Accounting Ratios	None

NN: Neutral Network, SOM: Self Organizing Map, RF: Random Forest, DT: Decision Tree, SVM: Support Vector Machine

Appendix B

Category of measurement	Variables	Studies
Efficiency	Asset Turnover Ratio	(Lokanan and Sharma, 2024)
	Total Sales/ Total Assets	(Nguyen *et al.*, 2023)
Liquidity	Current ratio	(Lokanan and Sharma, 2024; Tran *et al.*, 2022)
	Quick Ratio	(Lokanan and Sharma, 2024; Tran *et al.*, 2022)
	Working Capital/Total Assets	(Nguyen *et al.*, 2024)
Leverage	Short-term Debt to Assets	(Lokanan and Sharma, 2024; Tran *et al.*, 2022)
	Long-term Debt to Asset	(Lokanan and Sharma, 2024; Tran *et al.*, 2022)
	Market Value of Equity/ Total debt	(Nguyen *et al.*, 2024)
Profitability	Retained Earnings/ Total Assets	(Nguyen *et al.*, 2024)
	Earnings Before Interest and Taxes/Total Assets	(Pham Vo Ninh *et al.*, 2018)
	Net Income/ Total Assets	(Nguyen *et al.*, 2024)
	Profit margin	(Lokanan and Sharma, 2024)
Cashflow	Operating Cash flow/ Current Liabilities	(Lin and Dong, 2018)
	Change in Cashflow	(Lokanan and Sharma, 2024)
	Financial Expenses/ Cashflow	(du Jardin, 2018)
Valuation	Book Value of Equity/ Total Liabilities	(Pham Vo Ninh *et al.*, 2018)
	Enterprise Value to Revenues	(Tran *et al.*, 2022)

Appendix C7: Summary of descriptive statistics for independent variables in 3 year pre-distress

Variable	Mean		Median		Std.Dev		Min		Max	
	D	ND	D	ND	D	ND	D	ND	D	ND
MVETL	1.45	6.06	0.57	2.28	5.39	15.83	0.00	0.00	90.93	310.62
BVETL	0.85	4.06	0.48	1.99	2.15	6.86	-0.98	-0.45	32.15	75.26
WCTA	-0.25	0.29	0.03	0.29	2.32	0.24	-41.13	-1.48	0.93	0.97
RETA	0.06	0.11	0.03	0.07	0.08	0.11	0.00	0.00	0.72	0.65
ASSETURN	1.09	1.17	0.79	0.89	1.07	1.33	-0.01	0.00	8.76	16.53
SUPPAY	18.14	684.34	6.22	11.55	112.02	13988.41	0.00	0.00	2449.55	354082.27
CR	1.22	3.37	1.07	1.98	1.76	4.48	0.01	0.11	29.00	57.45
STDTA	0.33	0.09	0.19	0.03	1.12	0.12	0.00	0.00	16.10	1.41
LTDTA	0.11	0.04	0.03	0.02	0.20	0.08	0.00	0.00	2.46	0.59
FL	3.84	1.77	2.76	1.50	23.42	0.94	-97.81	-1.24	501.60	11.35
EBITIC	6.60	3930.44	1.46	5662.51	177.52	85537.29	-2317.34	-3188.85	2848.13	1922187.0
EBITTA	0.03	0.07	0.03	0.05	0.09	0.10	-0.27	-0.78	1.07	0.61
OCFTL	0.08	0.32	0.05	0.25	0.33	1.58	-2.50	-28.08	3.02	8.46
OCFCAPEX	-14.17	-5.15	-0.72	-0.54	664.17	245.89	-137.19	-4191.67	3434.81	3436.15
BPE	79.40	70.23	13.39	13.49	449.58	336.08	-1578.57	-898.67	8015.11	245315.45
EVR	9.44	385.78	0.39	1.00	155.65	9600.03	-1283.19	-0.36	3046.97	245315.45

D: Financial Distress , ND: Non Financial Distress

Appendix D 8: Summary of descriptive statistics for independent variables in 2 year Pevr re-distress

Variable	Mean		Median		Std.Dev		Min		Max	
	D	ND	D	ND	D	ND	D	ND	D	ND
MVETL	1.50	16.82	0.35	1.84	8.42	272.15	0.00	0.00	151.85	6963.24
BVETL	0.85	4.39	0.48	2.02	2.44	8.07	-0.98	0.11	33.42	98.34
WCTA	-0.30	0.32	0.04	0.31	2.89	0.22	-56.33	-0.36	0.96	0.97
RETA	0.05	0.10	0.03	0.07	0.06	0.11	0.00	0.00	0.38	0.73
ASSETURN	1.14	1.17	0.79	0.86	1.35	1.31	-0.01	-0.05	12.40	12.63
SUPPAY	19.48	651.57	6.07	11.96	84.61	13645.66	0.00	0.00	1275.71	345909.66
CR	1.19	3.65	1.07	2.04	1.86	5.96	0.01	0.03	38.26	107.39
STDTA	0.34	0.09	0.19	0.03	1.24	0.11	0.00	0.00	20.52	0.68
LTDTA	0.11	0.03	0.03	0.00	0.22	0.07	0.00	0.00	2.78	0.69
FL	-0.32	4.34	2.76	1.50	82.89	65.86	-1813.16	1.01	257.88	1690.88
EBITIC	-3.13	183.93	1.63	6.66	128.56	1992.66	-2201.39	-8601.98	903.87	34252.57
EBITTA	1.50	16.82	0.35	1.84	8.42	272.15	0.00	0.00	151.85	6963.24
OCFTL	0.85	4.39	0.48	2.02	2.44	8.07	-0.98	0.11	33.42	98.34
OCFCAPEX	-0.30	0.32	0.04	0.31	2.89	0.22	-56.33	-0.36	0.96	0.97
BPE	0.05	0.10	0.03	0.07	0.06	0.11	0.00	0.00	0.38	0.73
EVR	14.96	4.76	0.73	0.76	187.99	66.49	-52.87	-342.06	3853.34	1629.40

D: Financial Distress , ND: Non Financial Distress

Appendix E 9: Summary of descriptive statistics for independent variables in 1 year pre-distress

Variable	Mean		Median		Std.Dev		Min		Max	
	D	ND	D	ND	D	ND	D	ND	D	ND
MVETL	1.07	9.75	0.36	2.03	8.01	99.61	0.0	0.0	180.9	2519.61
BVETL	0.6	7.63	0.46	2.15	1.17	65.16	-0.99	0.12	20.92	1607.05
WCTA	-0.45	0.35	0.04	0.33	4.16	0.22	-78.81	-0.24	0.95	0.97
RETA	0.05	0.1	0.03	0.07	0.07	0.11	0.0	0.0	0.56	0.86
ASSETURN	1.37	1.31	0.88	0.93	2.54	1.65	-0.01	0.0	42.21	15.18
SUPPAY	21.44	664.83	6.76	13.02	85.43	15277.69	-0.05	0.0	1472.51	390045.05
CR	1.05	4.41	1.06	2.21	1.03	13.54	0.01	0.09	21.78	297.02
STDTA	0.39	0.08	0.21	0.03	1.62	0.11	0.0	0.0	26.31	0.73
LTDTA	0.12	0.03	0.03	0.0	0.34	0.06	0.0	0.0	6.11	0.42
FL	1.7	1.9	2.57	1.48	46.23	4.18	-1028.07	-20.1	103.86	95.13
EBITIC	1.38	235.96	1.45	1.94	111.72	3203.82	-1449.22	-6259.16	1505.45	60568.34
EBITTA	0.02	0.08	0.03	0.06	0.1	0.12	-1.07	-0.24	0.43	1.14
OCFTL	0.0	0.03	0.03	0.18	0.64	7.3	-8.95	-183.06	1.27	7.39
OCFCAPEX	4.77	-25.72	-0.14	0.39	184.79	680.18	-2139.93	-10996.64	2137.9	2160.71
BPE	33.72	203.14	8.03	9.09	266.26	4341.61	-2155.32	-2535.34	4067.09	110722.57
EVR	39.22	6.28	0.69	0.75	809.33	72.80	-102.23	-0.31	18252.28	1338.27

D: Financial Distress , ND: Non Financial Distress

Appendix F : Correlation Matrix for independent variables in 2020

	MVETL	BVETL	WCTA	RETA	ASSETURN	SUPPAY	CR	STDTA	LTDTA	FL	EBITIC	OCFTL	EBITTA	OCFCAPEX	BPE	EVR
MVETL																
BVETL	0.79															
WCTA	0.46	0.56														
RETA	0.12	0.16	0.38													
ASSETURN	-0.11	-0.14	0.13	0.30												
SUPPAY	0.29	0.36	0.25	0.20	0.30											
CR	0.63	0.79	0.82	0.25	-0.07	0.30										
STDTA	-0.45	-0.54	-0.42	-0.27	0.17	-0.13	-0.51									
LTDTA	-0.26	-0.32	-0.40	-0.23	-0.28	-0.15	-0.27	0.05								
FL	-0.54	-0.66	-0.43	-0.21	0.08	-0.26	-0.50	0.45	0.24							
EBITIC	0.04	0.02	0.02	0.07	0.04	-0.01	0.01	-0.04	-0.02	-0.03						
OCFTL	0.40	0.42	0.30	0.29	0.20	0.29	0.35	-0.31	-0.15	-0.36	0.06					
EBITTA	0.16	0.05	0.16	0.22	0.34	0.19	0.08	-0.06	0.00	-0.11	0.04	0.44				
OCFCAPEX	0.05	0.07	0.05	-0.06	-0.09	-0.02	0.08	-0.05	-0.10	-0.03	-0.01	-0.08	-0.16			
BPE	0.04	-0.01	0.02	-0.11	-0.09	-0.06	0.01	0.06	0.05	0.06	-0.01	-0.10	-0.23	-0.03		
EVR	0.37	0.20	-0.08	-0.26	-0.59	-0.08	0.13	-0.16	0.27	-0.16	-0.02	-0.08	-0.16	0.11	0.14	

Appendix G : Correlation Matrix for independent variables in 2021

	MVETL	BVETL	WCTA	RETA	ASSETURN	SUPPAY	CR	STDTA	LTDTA	FL	EBITIC	EBITTA	OCFTL	OCFCAPEX	BPE	EVR
MVETL																
BVETL	0.76															
WCTA	0.41	0.56														
RETA	0.15	0.16	0.33													
ASSETURN	-0.13	-0.20	0.11	0.20												
SUPPAY	0.30	0.33	0.24	0.15	0.27											
CR	0.60	0.80	0.80	0.20	-0.12	0.30										
STDTA	-0.45	-0.54	-0.40	-0.29	0.20	-0.13	-0.51									
LTDTA	-0.23	-0.30	-0.37	-0.18	-0.24	-0.13	-0.26	0.06								
FL	-0.51	-0.65	-0.41	-0.17	0.13	-0.25	-0.51	0.45	0.24							
EBITIC	0.14	0.14	0.12	0.11	0.03	-0.00	0.16	-0.09	-0.06	-0.07						
EBITTA	0.16	0.06	0.15	0.20	0.35	0.21	0.06	-0.05	0.06	-0.13	0.12					
OCFTL	0.31	0.29	0.09	0.26	-0.00	0.18	0.18	-0.31	-0.04	-0.25	0.10	0.43				
OCFCAPEX	0.04	0.06	0.01	-0.03	-0.13	-0.04	0.05	-0.03	-0.10	-0.08	-0.02	-0.19	-0.05			
BPE	0.10	0.08	0.09	-0.03	-0.06	0.02	0.11	-0.08	-0.01	-0.01	-0.02	-0.19	-0.06	-0.02		
EVR	0.32	0.19	-0.13	-0.25	-0.61	-0.10	0.11	-0.13	0.27	-0.17	-0.03	-0.20	-0.04	0.18	0.09	

Appendix H : Correlation Matrix for independent variables in 2022

	MVETL	BVETL	WCTA	RETA	ASSETURN	SUPPAY	CR	STDTA	LTDTA	FL	EBITIC	EBITTA	OCFTL	OCFCAPEX	BPE	EVR
MVETL																
BVETL	0.80															
WCTA	0.44	0.56														
RETA	0.18	0.14	0.30													
ASSETURN	-0.10	-0.18	0.10	0.30												
SUPPAY	0.32	0.34	0.22	0.14	0.29											
CR	0.65	0.80	0.80	0.20	-0.11	0.29										
STDTA	-0.45	-0.53	-0.42	-0.28	0.16	-0.13	-0.51									
LTDTA	-0.24	-0.29	-0.36	-0.18	-0.25	-0.13	-0.25	0.05								
FL	-0.43	-0.50	-0.36	-0.16	0.06	-0.23	-0.43	0.34	0.16							
EBITIC	0.09	0.10	0.08	0.07	0.04	-0.00	0.09	-0.07	-0.04	-0.05						
EBITTA	0.23	0.11	0.17	0.25	0.29	0.21	0.11	-0.15	0.06	-0.17	0.09					
OCFTL	0.38	0.34	0.17	0.26	0.07	0.21	0.26	-0.30	-0.06	-0.33	0.05	0.40				
OCFCAPEX	0.04	-0.00	-0.04	-0.03	-0.08	0.02	-0.04	0.01	0.01	0.03	-0.01	0.01	-0.01			
BPE	0.03	0.03	0.06	-0.04	-0.05	-0.04	0.05	-0.02	-0.01	0.03	-0.02	-0.19	-0.07	-0.04		
EVR	0.30	0.20	-0.08	-0.28	-0.62	-0.11	0.14	-0.14	0.25	-0.11	-0.02	-0.14	0.01	0.09	0.03	