Application of machine learning in assessing financial risk of listed companies on the Vietnam stock market

By QNU Journal of Science

WORD COUNT	4106	TIME SUBMITTED	17-MAR-2025 09:58AM
		PAPER ID	115186566

Ứng dụng học máy trong đánh giá rủi ro tài chính của các công ty niêm yết trên thị trường chứng khoán Việt Nam

TÓM TẮT

Quản lý rủi ro tài chính là điều cần thiết đối với các doanh nghiệp vì nó giúp ngăn ngừa tổn thất và tối đa hóa lợi nhuận. Do quá trình này phụ thuộc nhiều vào việc ra quyết định dựa trên dữ liệu, học máy mang lại tiềm năng phát triển các phương pháp và công nghệ sáng tạo. Trong bài báo này, chúng tôi so sánh khả năng dự đoán của các mô hình học máy khác nhau và sử dụng phương pháp LIME để diễn giải cách chúng đưa ra quyết định. Dữ liệu được thu thập từ báo cáo tài chính của các công ty niêm yết từ năm 2009 đến năm 2023. Kết quả cho thấy Gradient Boosting và Random Forest đạt hiệu suất tốt nhất. Thêm vào đó, trọng số LIME chỉ ra rằng các yếu tố ảnh hưởng nhiều nhất đến dự đoán của các mô hình là tỷ lệ thanh khoản hiện hành, tỷ suất lợi nhuận trên tài sản, tỷ lệ nợ và tỷ lệ nợ trên vốn chủ sở hữu.

Từ khóa: *Rủi ro tài chính, công ty niêm yết, mô hình học máy, phương pháp LIME.*

**Tác giả liên hệ chính.*
Email: bachathanh1@gmail.com

Application of machine learning in assessing financial risk of listed companies on the Vietnam stock market

ABSTRACT

Financial risk management is essential for businesses as it helps prevent losses and maximize profits. Since this process depends heavily on data-driven decision-making, machine learning offers a promising avenue for developing innovative methods and technologies. In this paper, we compare the predictive capabilities of various machine learning models and use the LIME method to interpret how they make decisions. Data was collected from the financial statements of listed companies from 2009 to 2023. The results show that Gradient Boosting and Random Forest achieved the best performance. Additionally, LIME weights indicate that the most influential factors affecting the models' predictions are the current ratio, return on assets, debt ratio, and debt-to-equity ratio.

Keywords: *Financial risk, listed companies, machine learning models, LIME method.*

1. INTRODUCTION

When a company is exposed to an event that can cause a shortfall in a targeted financial measure or value, this is financial risk.[1] The financial measure or value could be earnings per share, return on equity, or cash flows. Financial risks include market risk, credit risk, market liquidity risk, operational risk, and legal risk. Financial risk assessment is critical for investors, regulators, and corporate managers to identify potential challenges and mitigate their impacts.

Financial risk is often associated with the risk of bankruptcy or insolvency of a business. Traditional methods of financial risk assessment often rely on expert judgment and statistical models. Experts can leverage their domain knowledge to identify potential risks, assess the impact of external factors, and interpret the results of statistical models. However, expert judgment can be subjective and prone to bias, particularly when dealing with complex financial scenarios. Numerous statistical models have been proposed, such as Z-score, S-score, O-score, X-score, H-score, B-score,…[2-7] In Vietnam, researchers have tested the Z-score model in forecasting corporate failure[8] and bankruptcy[9], applied the B-score in analyzing factors influencing financial risk[10], compared various models in measuring financial distress[11],… Statistical models are straightforward in design, offer strong explanatory power, and require relatively short training time. However, they depend on numerous strict assumptions that are often unrealistic in practical scenarios, such as the presence of linear relationships, homogeneity of variances, and independence between variables. Violation of these assumptions can reduce the predictive power of statistical methods.[12]

In recent years, machine learning (ML) has emerged as a powerful tool for overcoming the limitations of traditional methods. ML algorithms can automatically learn complex patterns from large datasets, without relying on strict assumptions. This makes them well-suited for financial risk assessment, where data is often noisy, incomplete, and high-dimensional. Algorithms such as support vector machine (SVM), decision tree, and artificial neural network are applied to enhance the efficiency of traditional methods in volatility forecasting, bankruptcy prediction, credit scoring,…[13-16] Ensemble learning and hybrid models have been widely studied in this field. Research suggests that random forest algorithms may surpass other single or hybrid classifiers.[17,18]

In this article, we will construct and compare the performance of several advanced machine learning models, such as SVM, neural networks, random forests, gradient boosting,…in forecasting the financial risks of listed companies on the Vietnamese stock market. Additionally, we also assess the importance of features using LIME to identify the key factors influencing financial risk and propose solutions to mitigate these risks.

2. METHODOLOGY

2.1. Data collection and preprocessing

In this study, we utilize data extracted from the financial statements of companies listed on the

*Correspoding author.
Email: bachathanh1@gmail.com

HOSE (Ho Chi Minh Stock Exchange), HNX (Hanoi Stock Exchange), and UPCOM (Unlisted Public Company Market). The data spans the period from 2009 to 2023 and includes balance sheets, income statements, and cash flow statements.

This study applies machine learning models to predict financial risk, specifically bankruptcy risk, formulated as a classification problem. To identify companies at risk, we utilize five widely recognized bankruptcy prediction models: the Altman Z-score, Springate S-score, Zmijewski X-score, Taffler Z-score, and Grover G-score (Table 1). A company is labeled as 1 (at risk) if the majority of the five models classify it as being at risk, and -1 otherwise. Regarding independent variables, based on several studies, we use 34 financial ratios as inputs for the machine learning models, as presented in Table 2. These ratios reflect various aspects of the company, such as liquidity, profitability, efficiency, and leverage.

Table 1. Bankruptcy prediction models for defining the target variable.

Model	Formula	Conclusion
Z-score (1968)	Z = 1.2Z1 + 1.4Z2 + 3.3Z3+ 0.6Z4 + 1.0Z5 Z1 = Working capital / Total assets Z2 = Retained earnings / Total assets Z3 = EBIT / Total assets Z4 = Market value of equity / Total liabilities Z5 = Sales / Total assets	$Z < 2.99$: $y = 1$ $Z \geq 2.99$: $y = -1$
S-score (1978)	S = 1.03S1 + 3.07S2 + 0.66S3 + 0.4S4 S1 = Working capital / Total assets S2 = EBIT / Total assets S3 = Profit before tax / Current liabilities S4 = Sales / Total assets	$S < 0.862$: $y = 1$ $S \geq 0.862$: $y = -1$
X-score (1984)	X = -4.336 – 4.513X1 + 5.679X2 – 0.004X3 X1 = Net income / Total assets X2 = Total liabilities / Total assets X3 = Current assets / Current liabilites	$X \geq 0$: $y = 1$ $X < 0$: $y = -1$
Taffler Z-score (1983)	T = 3.31 + 12.18T1 + 2.50T2 – 10.68T3 + 0.029T4 T1 = Profit before tax / Current liabilities T2 = Current assets / Total liabilities T3 = Current liabilities / Total assets T4 = No-credit interval	$T \leq 0.3$: $y = 1$ $T > 0.3$: $y = -1$
G-score (2001)	G = 1.6505G1 + 3.404G2 – 0.016G3 + 0.057 G1 = Working capital / Total asssets G2 = EBIT / Total assets G3 = ROA	$G \leq 0.01$: $y = 1$ $G > 0.01$: $y = -1$

Table 2. Financial ratios (features) for assessing financial risk.

Symbol	Ratio name	Symbol	Ratio name
X1	Price-to-earnings ratio	X18	EV-to-EBIT ratio
X2	Price-to-sale ratio	X19	Price-to-operating- cash-flow ratio
X3	Price-to-book ratio	X20	Debt ratio
X4	Earnings per share	X21	Price-to-cash-flow ratio
X5	Return on equity	X22	Book value per share
X6	Return on assets	X23	Cash ratio
X7	Return on invested capital	X24	Return on capital employed
X8	Operating margin	X25	Return on sales
X9	Gross margin	X26	Cash return on invested capital

X10	Net margin	X27	Cash return on equity
X11	EBIT margin	X28	Cash return on assets
X12	Current ratio	X29	Free cash flow margin
X13	Quick ratio	X30	Operating cash flow margin
X14	Debt-to-equity ratio	X31	Total asset turnover ratio
X15	Operating cash flow ratio	X32	Equity ratio
X16	EV-to-EBITDA ratio	X33	Fixed asset turnover ratio
X17	EV-to-sales ratio	X34	Receivables turnover ratio

The dataset consists of 2614 observations, including 557 observations with $y = 1$ and 2057 observations with $y = -1$. Before performing preprocessing steps, the data is split into training and testing sets at an 8:2 ratio to prevent data leakage. Data leakage in machine learning occurs when a model uses information during training that wouldn't be available at the time of prediction. Leakage causes a predictive model to look accurate until deployed in its use case; then, it will yield inaccurate results, leading to poor decision-making and false insights.[19] The dataset is then cleaned by removing outliers and imputing missing values.

2.2. Dimensionality reduction

The dimensionality reduction involves reducing the number of variables to be used for efficient modeling. Dimensionality reduction can help to reduce the complexity of the model and improve its generalization performance. There are two main approaches to dimensionality reduction: feature selection and feature extraction. Feature selection involves selecting a subset of the original features that are most relevant to the problem at hand. Feature extraction involves creating new features by combining or transforming the original features.

Here, we will use the feature selection to retain the original meaning of the variables in the dataset. Our data has numerical attributes, and the target variable is categorical, so we will use the ANOVA F-test technique.[20] ANOVA, or "analysis of variance" a parametric statistical test used to determine whether the means of two or more data samples (typically three or more) come from the same distribution. It is a type of F-test, which refers to statistical tests that compare variance values, such as the variance between different samples or the explained and unexplained variance in a test like ANOVA. This method is particularly useful when one variable is numerical and the other is categorical, such as numerical input features and a categorical target variable in classification tasks. The results of ANOVA can be applied in feature selection by identifying and removing features that are independent of the target variable, helping to refine the dataset for better model performance.

2.3. Machine learning models to predict financial risk

In this study, we implement and compare the effectiveness of statistical and machine learning models, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost), Gradient Boosting, and Multi-layer Perceptron (MLP).

2.3.1. Logistic Regression

Logistic regression is a widely used statistical method for predicting outcomes in scenarios where the dependent variable is binary.[21] In this study, it is applied to determine financial risk status. The model produces an output P_n, which represents the probability of being at risk based on the input variables $\boldsymbol{X}$. This probability is derived using Equation (1).

$$P_n(y = 1|\boldsymbol{X}) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \quad (1)$$

Logistic regression is commonly employed as a benchmark in research for evaluating the performance of other forecasting methods. Its primary strength lies in the simplicity and clarity of its results, making them accessible and easy to interpret for most users. This high level of interpretability makes logistic regression a popular choice in practical applications, particularly within financial institutions.

2.3.2. Support Vector Machine

Support Vector Machine (SVM) is a robust machine learning algorithm designed for both classification and regression tasks.[22] In this study, it is employed to classify data points into distinct categories based on input features $\boldsymbol{X}$. The model constructs an optimal hyperplane that maximizes the margin between support vectors. The classification process will take place according to Equation (2):

$$y = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b) \quad (2)$$

SVM is particularly effective in handling high-dimensional data and is often combined with kernel functions to address non-linear problems. Its main advantage lies in its ability to generalize well, even with limited data, making it a standard choice in applications like image classification, bioinformatics, and text categorization.

2.3.3. Random Forest

Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks. It builds multiple decision trees during training and combines their outputs to improve accuracy and reduce overfitting.[23] Each tree is trained on a random subset of data, and only a random subset of features is considered for splitting at each node, enhancing diversity among the trees. The final prediction is made through majority voting (for classification) or averaging (for regression). Known for its robustness and ability to handle high-dimensional, non-linear data, Random Forest is widely applied in areas like financial risk assessment, medical diagnosis, and image classification.

2.3.4. Adaptive Boosting

Boosting algorithms work on the idea of first building a model on the training dataset and then building a second model to correct the faults in the first model. This technique is repeated until the mistakes are reduced and the dataset is accurately predicted. Boosting algorithms function similarly in that they combine numerous models (weak learners) to produce the final result (strong learners).

AdaBoost works by initially assigning equal weights to all samples in the training dataset.[24] The algorithm then iterates for a predefined number of iterations or until a stopping criterion is met. In each iteration, a weak classifier $\hat{f}_i$ (e.g., a one-level decision tree) is trained on the data. The weights of the samples are updated, giving higher weights to misclassified examples to focus more on them in subsequent iterations. The weak classifiers are evaluated based on their errors, with lower-error classifiers receiving higher weights. The sample weights are then normalized to sum up to 1. The final prediction is made by combining the predictions of all p weak classifiers using a weighted majority vote:

$$\hat{f}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{p} \alpha_i \hat{f}_i(\boldsymbol{x})\right) \quad (3)$$

This process repeats until the specified number of iterations is completed or the stopping criterion is satisfied.

2.3.5. Gradient Boosting

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient.[25] The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

In contrast to AdaBoost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels. There is a popular technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

2.3.6. Multi-layer Perceptron (MLP)

Multi-layer Perceptron is the most known and most frequently used type of neural network architecture that consists of three main types of layers: an input layer, one or more hidden layers, and an output layer.[26] Each neuron in a hidden layer processes a weighted sum of its inputs, followed by a non-linear activation function such as Rectified Linear Unit (ReLU), Sigmoid, or Hyperbolic Tangent (Tanh).

During training, MLP utilizes a two-step learning process: forward propagation and backpropagation. In forward propagation, the output of a neuron is computed as follows:

$$\boldsymbol{z}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)} \quad (4)$$

$$\boldsymbol{a}^{(l)} = f\left(\boldsymbol{z}^{(l)}\right) \quad (5)$$

where $\boldsymbol{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$ are the weight matrix and bias vector for layer l, $\boldsymbol{x}^{(l-1)}$ is the input from the previous layer, and $f(.)$ is the activation function. The backpropagation algorithm then updates the network's weights by computing gradients of the loss function with respect to the weights using the chain rule. The gradient descent optimization technique, often with variations such as Stochastic Gradient Descent (SGD) or Adam, is applied to minimize the loss iteratively.

MLP is widely used in classification and regression tasks due to its ability to learn complex patterns in data. It serves as a foundation for more advanced deep learning models and is particularly effective in applications such as image recognition, speech processing, and time series prediction.

2.3. Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.[27]

The LIME explainability model belongs to what is often referred to as "surrogate models". The creation of this surrogate model is done in a step wise process. First, the LIME algorithm creates a new proxy dataset by making slight permutations to the feature values of the available dataset (that is: the dataset that was used to train the black-box model). Next, each of these samples is assigned a weight that is proportional to its similarity with respect to the instance we are trying to explain. At last, a surrogate machine learning model – which is an explainable model such as a decision tree classifier/regressor or logistic regression model — is trained on the weighted proxy dataset. The learned model should be a good approximation of the machine learning model predictions locally, but it does not have to be a good global approximation. This kind of accuracy is also called local fidelity. The explanation produced by LIME is obtained by the following:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (6)$$

The explanation model for instance x is the model g (e.g. linear regression model) that minimizes loss $\mathcal{L}$ (e.g. mean squared error), which measures how close the explanation is to the prediction of the original model f (e.g. an AdaBoost model), while the model complexity $\Omega(g)$ is kept low. G is the family of possible explanations, for example all possible linear regression models. The proximity measure π_x defines how large the neighborhood around instance x is that we consider for the explanation.

3. RESULTS AND DISCUSSION

3.1. Dimensionality reduction results

Figure 1 illustrates the F-statistics of 34 features when performing ANOVA. X14 (debt-to-equity ratio) appears to be the most relevant, and 16 out of 34 features have significantly higher scores compared to the rest. We will retain these 16 features and remove the other 18 to proceed with building machine learning model.

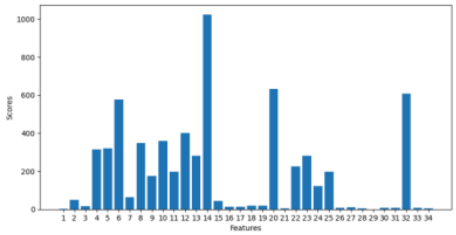

Figure 1. ANOVA F-test result.

3.2. Performance of machine learning models

When training a machine learning model, we fit the model's parameters. However, before the model begins learning, certain parameters are pre-set—these are called hyperparameters. We can improve the model's performance by tuning these hyperparameters. There are several hyperparameter tuning methods, such as grid search, random search, and Bayesian optimization. Among them, grid search is widely used. Grid search constructs a grid of all possible hyperparameter combinations. It then trains and evaluates the model with each combination, selecting the best performer. This thorough exploration of the hyperparameter space ensures no stone is left unturned. Moreover, grid search is typically used with cross-validation, specifically k-fold cross-validation. Here, the training set is divided into k parts. In each iteration, $k-1$ parts are used to train the model, while the remaining part is used for validation. The best set of hyperparameters is the one that yields the highest average performance. Finally, the models with the optimal set of hyperparameters are tested on the test set using various metrics.

Table 3 presents the hyper-parameter settings and the evaluation of the models on different metric

Table 3. Performance of the models on test set.

Models	Hyper-parameter settings	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	C=1, max_iter=300, penalty='1', solver= 'saga'	0.9331	0.8861	0.8952	0.8901	0.9693
SVM	C=1, degree=2, gamma='scale', kernel='rbf'	0.8642	0.7799	0.8729	0.8097	0.9534
Random Forest	bootstrap=False, max_depth=10, max_features='sqrt', min_samples_split=20, n_estimators=100	0.9484	0.9133	0.9166	0.9149	0.9836
AdaBoost	learning_rate=1, n_estimators=500	0.9331	0.8904	0.8873	0.8888	0.9780
Gradient Boosting	learning_rate=0.5, loss='log_loss', max_depth=7, max_features='sqrt', min_samples_split=20, n_estimators=100	0.9579	0.9276	0.9344	0.9309	0.9870
MLP	activation='relu', alpha=0.01, hidden_layer_sizes=(100,), learning_rate='adaptive', solver='adam'	0.9312	0.8785	0.9020	0.8896	0.9788

Gradient Boosting achieved the best performance across all metrics, indicating high predictive accuracy and a good balance between precision and recall. Random Forest ranked second with high accuracy and AUC, demonstrating strong and consistent classification ability. MLP also showed good results across all metrics, particularly in AUC.

AdaBoost and Logistic Regression had similar performance with accuracy but showed lower precision and recall compared to Gradient Boosting and Random Forest. SVM had the lowest performance across all metrics, particularly in precision and F1 score, indicating difficulties in accurate classification and balancing precision and recall.

Overall, Gradient Boosting is the most suitable model for this problem, followed by Random Forest and MLP, while SVM performed the worst.

3.3. Interpretations of results

We used LIME to interpret the two best-performing models: Gradient Boosting and Random Forest. A random instance from the test set was selected to generate a local explanation for this specific instance (Figure 2).

The chosen instance has a true label of $y = -1$, indicating no risk. Both models identified features X12 and X24 as the most influential. Specifically, X12 contributes to the model's prediction of $y = -1$, while X24 influences the prediction in the opposite direction. For Gradient Boosting, the impact of features decreases noticeably from top to bottom, highlighting the model's tendency to focus on the most important features. In contrast, Random Forest distributes influence more evenly across features, reflecting its nature of aggregating predictions from multiple independent decision trees.

Local explanations are valuable for understanding the reasoning behind individual predictions. However, analyzing a single instance does not provide a comprehensive understanding of the model's overall behavior. To gain deeper insights into the model's decision-making process, we can aggregate local explanations across multiple predictions. Specifically, by combining the LIME weights of numerous instances and visualizing them through various types of charts, we can better capture the model's general patterns and feature importance.

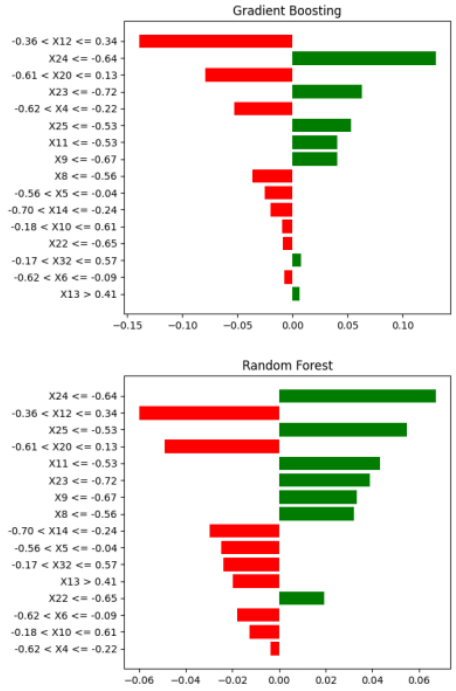

Figure 2. Local explainations of Gradient Boosting and Random Forest.

The first aggregation can help us understand which of the features are most important. Features with either high positive or negative LIME weights had a larger impact on a prediction. For each feature, we take the absolute mean of all the LIME weights. Features with large mean weights have, in general, made large contributions to the predictions. Figure 3 shows the average weights of the features in the two models. It can be observed that the important features are relatively similar across both models. These features are the current ratio (X12), return on assets (X6), debt ratio (X20), and debt-to-equity ratio (X14).

Next, we will examine the trends of these important features by plotting the LIME weights against their values. A high LIME weight indicates that, for a specific prediction, the value of this feature increases the probability of predicting a "At risk" case ($y = 1$).

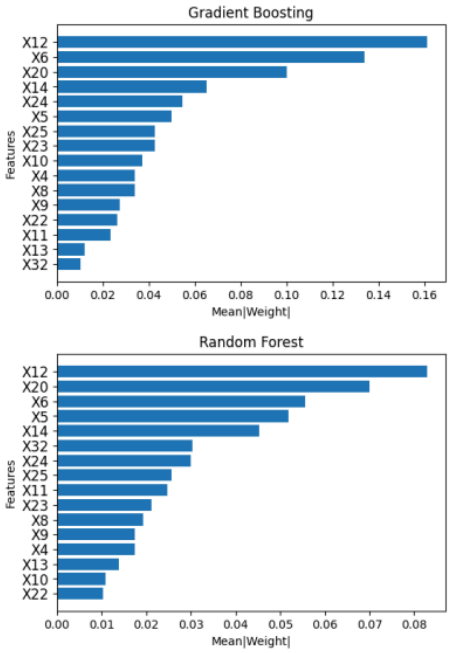

Figure 3. Absolute mean of LIME weights.

Figure 4 illustrates that as the values of X12 and X6 rise, their LIME weights shift from positive to negative. X12, representing the current ratio, assesses a company's short-term liquidity. A low current ratio suggests potential liquidity problems, which increase financial risk and result in a positive LIME weight. In contrast, a high current ratio indicates a stronger ability to meet debt obligations, reducing financial risk and producing a negative LIME weight. This negative weight decreases the probability of being classified as risky ($y = 1$). Meanwhile, X6, which measures return on assets (ROA), reflects how efficiently a company generates profit from its assets. A low ROA indicates weak profitability and higher financial risk, leading to a positive LIME weight. Conversely, a high ROA signifies effective asset management and lower risk, resulting in a negative LIME weight.

On the other hand, the LIME weights for X20 and X14 increase as their values grow. X20, the debt ratio, indicates the proportion of a company's assets financed through debt. A high debt ratio suggests significant reliance on borrowed funds, which raises financial leverage and risk due to fixed interest obligations. Similarly, X14, the debt-to-equity ratio, compares total debt to shareholders' equity. A high value for X14 indicates a greater

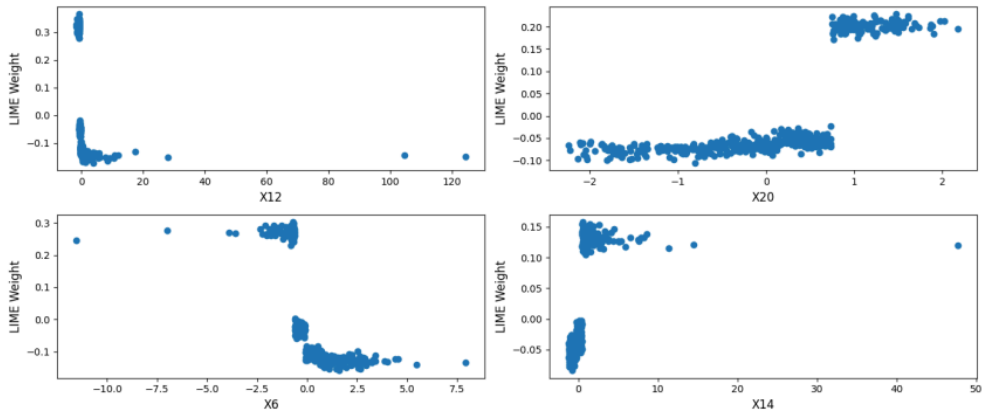

Figure 4. Feature trends for the four most important features.

dependence on debt compared to equity, leading to increased financial burden and risk.

3. CONCLUSIONS

In this study, we developed and compared advanced machine learning models to predict the financial risk of companies listed on the Vietnamese stock market. Based on financial ratios, various models were constructed, hyperparameters were optimized, and evaluations were conducted using different metrics. The two best-performing models were Gradient Boosting and Random Forest, achieving over 94% accuracy and more than 91% recall. This demonstrates the superiority of ensemble learning methods over single models. Furthermore, the LIME method was utilized to explain the models' predictions and the influence of different features on their decisions. The results indicate that to reduce financial risk, businesses should improve their current ratio (X12) by efficiently managing inventory and accelerating receivables collection, thereby reducing the likelihood of liquidity issues. Additionally, enhancing return on assets (ROA - X6) through optimized production processes can lower financial risk. Companies should also closely monitor the debt ratio (X20) and debt-to-equity ratio (X14) by avoiding excessive borrowing and increasing equity financing to reduce interest burdens. Moreover, diversifying funding sources by balancing debt and equity financing will optimize the capital structure and minimize financial risk in the long term. The findings of this study provide a foundation for businesses to manage risks more effectively, make safer business decisions, and optimize their strategies.

Application of machine learning in assessing financial risk of listed companies on the Vietnam stock market

ORIGINALITY REPORT

39%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.mdpi.com Internet	158 words — 4%
2	www.geeksforgeeks.org Internet	152 words — 3%
3	medium.com Internet	127 words — 3%
4	almabetter.com Internet	63 words — 1%
5	fastercapital.com Internet	62 words — 1%
6	ebin.pub Internet	57 words — 1%
7	epdf.pub Internet	57 words — 1%
8	link.springer.com Internet	55 words — 1%
9	www.cdn.geeksforgeeks.org Internet	54 words — 1%
10	www.ibm.com Internet	52 words — 1%

11	keylabs.ai Internet	37 words — 1%
12	journal.fsv.cuni.cz Internet	35 words — 1%
13	pavanrajkumar.github.io Internet	33 words — 1%
14	Bertrand Iooss, Ron Kenett, Piercesare Secchi. "Chapter 1 Different Views of Interpretability", Springer Science and Business Media LLC, 2022 Crossref	29 words — 1%
15	123docz.net Internet	27 words — 1%
16	Thaung Myint Tun, Lay Lay Myat, Zaw Tun. "Student Dropout Rate Analysis using Education Data of Rakhine State", 2023 IEEE Conference on Computer Applications (ICCA), 2023 Crossref	24 words — 1%
17	Xiaoxuan Zhang, Da Wang, Huchao Ma, Saina Dong, Zhiyu Wang, Zhenlei Wang. "Application of Machine Learning to Research on Trace Elemental Characteristics of Metal Sulfides in Se-Te Bearing Deposits", Minerals, 2024 Crossref	24 words — 1%
18	arxiv.org Internet	22 words — 1%
19	img1.wsimg.com Internet	22 words — 1%
20	Saravanan Parthasarathy, Vaishnavi Jayaraman, Jane Preetha Princy R. "Predicting Heart Failure using SMOTE-ENN-XGBoost", 2023 International Conference on	20 words — < 1%

Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 2023
Crossref

21	P.V. Mohanan. "Artificial Intelligence and Biological Sciences", CRC Press, 2025 Publications	19 words — < 1%
22	sciencen.org Internet	19 words — < 1%
23	www.frontiersin.org Internet	19 words — < 1%
24	"Intelligent Robots and Drones for Precision Agriculture", Springer Science and Business Media LLC, 2024 Crossref	18 words — < 1%
25	toc.123doc.org Internet	18 words — < 1%
26	www.researchgate.net Internet	18 words — < 1%
27	Jon Andreas Mortensen, Martin Efremov Mollov, Ayan Chatterjee, Debasish Ghose, Frank Y. Li. "Multi-Class Stress Detection through Heart Rate Variability: A Deep Neural Network based Study", IEEE Access, 2023 Crossref	17 words — < 1%
28	dokumen.pub Internet	17 words — < 1%
29	udspace.udel.edu Internet	17 words — < 1%
30	www.gridlex.com Internet	17 words — < 1%

31	D.A. Langford, A. Retik. "The Organization and Management of Construction - Shaping Theory and Practice", Routledge, 2012 Publications	16 words — < 1%
32	Tao Hou, Jing Li. "Application of mask R-CNN for building detection in UAV remote sensing images", Heliyon, 2024 Crossref	16 words — < 1%
33	Trương Xuân Nam, Nguyễn Thanh Tùng. "DEEP LEARNING: ỨNG DỤNG CHO DỰ BÁO LƯU LƯỢNG NƯỚC ĐẾN HỒ CHỨA HÒA BÌNH", FAIR - NGHIÊN CỨU CƠ BẢN VÀ ỨNG DỤNG CÔNG NGHỆ THÔNG TIN - 2016, 2017 Crossref	16 words — < 1%
34	Brad Boehmke, Brandon Greenwell. "Hands-On Machine Learning with R", CRC Press, 2019 Publications	15 words — < 1%
35	www.baeldung.com Internet	15 words — < 1%
36	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publications	14 words — < 1%
37	Nan Cao, Michael C.P. Sing. "Workforce forecasting in the building maintenance and repair work: Evaluating machine learning and LSTM models", Journal of Building Engineering, 2024 Crossref	14 words — < 1%
38	VNUA Publications	14 words — < 1%
39	bussecon.com Internet	14 words — < 1%

40	Biswajit Jena, Sanjay Saxena, Sudip Paul. "Machine Learning for Neurodegenerative Disorders - Advancements and Applications", CRC Press, 2025 Publications	12 words — < 1%
41	Kim Long Tran, Hoang Anh Le, Cap Phu Lieu, Duc Trung Nguyen. "Machine Learning to Forecast Financial Bubbles in Stock Markets: Evidence from Vietnam", International Journal of Financial Studies, 2023 Crossref	12 words — < 1%
42	eitca.org Internet	12 words — < 1%
43	repository.lib.ncsu.edu Internet	12 words — < 1%
44	text.123docz.net Internet	12 words — < 1%
45	core.ac.uk Internet	11 words — < 1%
46	journal.uc.ac.id Internet	11 words — < 1%
47	repository.unibos.ac.id Internet	11 words — < 1%
48	tapchi.ftu.edu.vn Internet	11 words — < 1%
49	www.teses.usp.br Internet	11 words — < 1%
50	"Deep Learning Theory and Applications", Springer Science and Business Media LLC, 2024 Crossref	10 words — < 1%

51	bmcpublichealth.biomedcentral.com Internet	10 words — < 1%
52	Thanasis Antamis, Charalampos-Rafail Medentzidis, Michael Skoumperdis, Thanasis Vafeiadis et al. "AI-supported Forecasting of Intermodal Freight Transportation Delivery Time", 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), 2021 Crossref	9 words — < 1%
53	Vũ Đình Tuấn, Hoàng Nhật Đức, Trần Xuân Linh. "DỰ BÁO XÓI MÒN ĐẤT DO MƯA GÂY RA Ở VÙNG ĐỒI NÚI VIỆT NAM BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY", KỶ YẾU HỘI THẢO CAREES 2019 NGHIÊN CỨU CƠ BẢN TRONG LĨNH VỰC KHOA HỌC TRÁI ĐẤT VÀ MÔI TRƯỜNG, 2019 Crossref	9 words — < 1%
54	icmai.in Internet	9 words — < 1%
55	luanvan.co Internet	9 words — < 1%
56	robots.net Internet	9 words — < 1%
57	Anurag Tiwari, Manuj Darbari. "Emerging Trends in Computer Science and Its Application - Proceedings of the International Conference on Advances in Emerging Trends in Computer Applications (ICAETC-2023) December 21–22, 2023, Lucknow, India", CRC Press, 2025 Publications	8 words — < 1%
58	He Lan, Shutian Wang, Wenfeng Zhang. "Predicting types of human-related maritime accidents with explanations using selective ensemble learning and SHAP method", Heliyon, 2024 Crossref	8 words — < 1%

59	caphesach.wordpress.com Internet	8 words — < 1%
60	entri.app Internet	8 words — < 1%
61	fwps.ftu.edu.vn Internet	8 words — < 1%
62	ijsra.net Internet	8 words — < 1%
63	iq.opengenus.org Internet	8 words — < 1%
64	repo.undiksha.ac.id Internet	8 words — < 1%
65	su-plus.strathmore.edu Internet	8 words — < 1%
66	vi.wikipedia.org Internet	8 words — < 1%
67	www.aitvn.asia Internet	8 words — < 1%
68	Muhammad Faraz Manzoor, Muhammad Shoaib Farooq, Adnan Abid. "Stylometry-driven framework for Urdu intrinsic plagiarism detection: a comprehensive analysis using machine learning, deep learning, and large language models", Neural Computing and Applications, 2025 Crossref	6 words — < 1%

EXCLUDE QUOTES OFF EXCLUDE SOURCES OFF

EXCLUDE BIBLIOGRAPHY OFF

EXCLUDE MATCHES OFF