ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

ΑΔΣ

Advances in Decision Sciences

Volume 29
Issue 4
December 2025

亞洲大學 ASIA UNIVERSITY

SCIENTIFIC & BUSINESS WORLD

Published by Asia University, Taiwan and Scientific and Business World

Forecasting ROA and ROE for Retail Companies in Vietnam by Using Machine Learning Techniques

Quang Hung Do

Fintech Lab, Posts and Telecommunications Institute of Technology,

Hanoi, Vietnam

***Corresponding author** Email: dqhung@ptit.edu.vn

ORCID: 0000-0002-8937-5102

Received: March 20, 2025; First Revision: August 11, 2025;

Last Revision: November 22, 2025; Accepted: December 9, 2025;

Published: December 11, 2025

Abstract

Purpose: This study aims to forecast the financial performance of Vietnamese retail companies by predicting both Return on Assets (ROA) and Return on Equity (ROE) by using artificial intelligence (AI) models, thereby enhancing predictive decision-making in the retail sector.

Design/methodology/approach: Financial statements from publicly listed retail firms covering the period 2010 to 2024, together with macroeconomic variables (CPI, exchange rate, gold price, VN-Index, oil prices), are combined. Three machine learning algorithms—Random Forest, XGBoost, and Multilayer Perceptron—are trained and tested by using an 80/20 split. Model accuracy is assessed by using RMSE, MAE, MAPE, Pearson's R, and Theil's U.

Findings: Random Forest achieved the lowest RMSE of 0.0926 and delivered the highest forecasting accuracy for both ROA and ROE, followed by XGBoost, while MLP performed less effectively. Decision tree–based ensemble models better capture non-linear relationships in financial data than neural networks in this context. Research limitations/implications – The study covers only listed retail firms and annual data, omitting private firms and intra-year fluctuations. The models identify associations, not causality.

Practical implications: From a Decision-Sciences perspective, the proposed AI-driven forecasting framework operates as a data-rich decision-support system that transforms multi-source information into actionable guidance on capital allocation, risk mitigation, and performance control.

Originality/value: This is the first study to jointly forecast both ROA and ROE for Vietnamese retailers using three complementary ML families, integrating firm-level and macro-financial variables, and providing reproducible benchmarks for AI-driven financial forecasting in emerging markets.

Keywords: Financial performance, forecasting, ROA, ROE, retail enterprises, Vietnam.

JEL Classifications: C01, C02, G17

1. Introduction

Financial performance is one of the key factors determining a firm's success and sustainability, particularly in the retail sector—a highly competitive industry that is significantly influenced by macroeconomic fluctuations. Two widely used indicators for assessing a company's financial performance are Return on Assets (ROA) and Return on Equity (ROE). The retail sector in Vietnam plays a crucial role in the national economy, contributing significantly to GDP growth, employment, and consumer spending. With a rapidly expanding middle class and increasing urbanization, Vietnam's retail market has witnessed substantial growth, attracting both domestic and foreign investments. However, this sector is also highly competitive and sensitive to macroeconomic fluctuations, including changes in inflation, exchange rates, and global commodity prices. Given these dynamics, accurately forecasting ROA and ROE is essential for retail companies to optimize their financial strategies, mitigate risks, and enhance decision-making processes. Effective financial forecasting enables firms to allocate resources efficiently, improve operational resilience, and maintain competitive advantages in an evolving economic landscape. ROA reflects a firm's ability to utilize its assets to generate profit, whereas ROE indicates the efficiency of equity utilization. Accurate forecasting of these indicators not only enables businesses to evaluate their financial strategies but also assists investors and managers in making informed decisions (Zafar & Yasin, 2025).

Currently, financial forecasting primarily relies on traditional methods such as Linear Regression, Time Series Models, and financial analysis based on accounting ratios. However, these approaches may struggle to capture the complexity and non-linearity of financial data, particularly in the context of a rapidly changing economy (Tutcu et al., 2024). The advancement of Artificial Intelligence (AI) and Machine Learning (ML) has introduced a new paradigm in financial forecasting, enhancing prediction accuracy and enabling the identification of hidden patterns within data. AI/ML models possess the capability to process large datasets and integrate multiple macroeconomic and firm-level financial factors, thereby improving forecasting precision.

This study aims to apply AI/ML techniques to develop predictive models for ROA and ROE in Vietnamese retail companies. The research utilizes financial statement data from retail firms spanning the period 2010–2024, combined with macroeconomic variables such as Consumer Price Index (CPI), exchange rates, gold prices, VN-Index, and oil prices. By comparing various AI/ML models, this study evaluates forecasting performance and identifies key factors influencing ROA and ROE in the retail sector.

This research contributes to financial forecasting in two main aspects. First, it provides a comprehensive assessment of AI/ML models in forecasting ROA and ROE by comparing methods such as Random Forest, Gradient Boosting, and Neural Networks. Second, it analyzes the significance of macroeconomic variables, offering valuable insights for businesses and investors regarding the key factors affecting the financial performance of retail firms in Vietnam.

This study is the first to (i) model ROA and ROE simultaneously for Vietnamese retailers with machine-learning techniques, (ii) fuse firm-level predictors with a rich set of macroeconomic indicators in one unified forecasting framework, and (iii) benchmark ensemble tree methods against a neural architecture using identical train–test splits. By doing so, we extend the fast-growing literature on AI-driven financial analytics (e.g., Đạt et al., 2025; He et al., 2023; Kayakus et al., 2023; Mihali & Niță, 2024; Nhật & Duy, 2024) to a data-sparse emerging economy and provide methodological guidance for scholars and practitioners. These novelties, together with the decade-long coverage from 2010–2024, differentiate our work decisively from earlier Vietnamese studies that considered either a single profitability metric or a restricted variable set.

By casting our predictive models as a decision-support tool that sharpens managerial choices under uncertainty, this study directly advances the central objective of Decision Sciences—enhancing the quality and speed of complex organisational decisions.

2. Background and Theoretical Framework

2.1. Theoretical foundations

The empirical analysis in this paper is grounded in two complementary strands of financial theory. First, the DuPont system (DuPont identity) decomposes firm profitability and links Return on Equity (ROE) and Return on Assets (ROA) to three broad drivers: profit margin (profitability), asset turnover (efficiency), and financial leverage (equity multiplier). This decomposition explains how operating performance, asset utilisation, and capital structure jointly determine observed returns (DuPont identity; see e.g., Healy & Palepu, 2001). In the context of retail firms, the DuPont decomposition motivates the inclusion of asset turnover and leverage indicators because retail profitability is typically driven by sales efficiency and the degree of gearing.

Second, corporate finance theory emphasises that capital structure (leverage), liquidity, and working capital management directly affect firm risk and expected returns (e.g., Healy & Palepu, 2001). Macroeconomic shocks (inflation, exchange rates, commodity prices) affect retail margins via input costs, demand shifts, and inventory revaluation. Together, these theoretical perspectives justify combining firm-level accounting ratios with macro-financial variables in a joint forecasting framework for ROA and ROE.

2.2. Empirical literature: forecasting ROA and ROE

A voluminous literature documents both parametric and machine-learning approaches to forecasting firm performance. Traditional econometric and accounting studies commonly use ratio analysis, panel regressions, and time-series models to explain ROA/ROE dynamics (Healy & Palepu, 2001; Hyndman & Athanasopoulos, 2018). In recent years, machine learning (ML) algorithms have been increasingly applied to firm-level forecasting tasks because of their ability to capture nonlinearities and high-order interactions (Breiman, 2001; Chen, 2025). Systematic reviews show rising adoption of ML for firm performance prediction and solvency tasks (e.g., machine-learning reviews).

Several applied studies forecast ROA/ROE using ensemble trees and neural networks. Kayakus et al. (2023) and Tutcu et al. (2024) compare tree-based methods with ANN in industry-specific settings and report that ensemble tree models often achieve superior out-of-sample performance in relatively small samples. Studies focused on emerging markets, including some recent Vietnam applications, provide evidence that Random Forest and related algorithms produce accurate profitability forecasts compared with simple linear baselines (Đạt et al., 2025; Pham & Le, 2024).

In addition to forecasting approaches based on machine learning, recent econometric literature has emphasized the risk of spurious regression and misleading inferences even when working with stationary or nearly stationary series (Cheng et al., 2021, 2022; Wong et al., 2024a, b; Wong & Yue, 2024). These works provide important cautionary evidence that conventional significance testing may generate unreliable results if the time-series properties of the data are not carefully examined. By drawing attention to spurious problems in financial modeling, this stream of research strengthens the methodological rigor required in profitability forecasting.

2.3. Research gap and contribution

Existing Vietnam-focused research tends to (i) examine either ROA or ROE rather than both jointly, (ii) rely on conventional regression frameworks or single algorithms, and (iii) study narrower sets of predictors. This study addresses these gaps by (i) jointly forecasting ROA and ROE in the retail sector, (ii) combining three complementary ML families (bagging, boosting, and neural networks), and (iii) integrating a richer set of macro-financial covariates with firm accounting data. This design is justified theoretically (DuPont and corporate finance channels) and offers practical value for managers and regulators in an emerging-market setting.

3. Data and Methodology

3.1. Data and Sample

Data sources and period: Financial statement variables for listed retail firms (Table 1) are collected from company annual reports and stock exchanges (HOSE, HNX, UPCOM). Macroeconomic variables (CPI, USD/VND, gold price, WTI, Brent, VN-Index) are taken from the General Statistics Office of Vietnam and public financial data providers (detailed sources listed in the variable table). The sample period is 2010–2024.

Table 1: List of Retail Companies Included in the Study

Company Code	Company Name	Stock Exchange
FRT	FPT Digital Retail Joint Stock Company	HOSE
PNJ	Phu Nhuan Jewelry Joint Stock Company	HOSE
MSN	Masan Group Corporation	HOSE
MWG	Mobile World Investment Corporation	HOSE
DGW	Digiworld Corporation	HOSE

BTT	Ben Thanh Trading & Service Joint Stock Company	HOSE
CMV	Ca Mau Trading Joint Stock Company	HOSE
CTF	City Auto Joint Stock Company	HOSE
HAX	Hang Xanh Automobile Service Joint Stock Company	HOSE
HTC	Hoc Mon Trading Joint Stock Company	HNX
TMC	Thu Duc Trading and Import-Export Joint Stock Company	HNX
MCH	Masan Consumer Corporation	UPCOM

Sample selection and final sample size: Firms were selected based on (i) SIC/industry classification as retail; (ii) availability of audited annual reports; and (iii) listed status during the sample period. After cleaning for missing/erroneous observations (see Section 3.3), the final working sample contains 168 firm-year observations.

Although the sample size of 168 firm-year observations is modest for machine learning, especially for a data-intensive model such as MLP, several steps were taken to mitigate the risk of overfitting. First, the MLP architecture was deliberately kept simple, with only two hidden layers (64 and 32 neurons) to balance model capacity and data availability. Second, early stopping with a patience of 10 epochs was applied during training to halt learning when validation performance ceased to improve. Third, hyperparameters for all models were tuned using grid search combined with k-fold cross-validation on the training set, ensuring that parameter choices generalised beyond the specific train–test split. These measures collectively reduce variance and help prevent the models—particularly MLP—from memorising noise rather than learning robust patterns. Table 2 summarises variable names, abbreviations, definitions, and data sources.

Table 2: Variables

Variable name	Abbreviation	Definition (calculation)	Data source
Return on Assets	ROA	Net Income / Average Total Assets	Company annual reports
Return on Equity	ROE	Net Income / Average Shareholders' Equity	Company annual reports
Total assets	TotalAssets	Book value of total assets (year-end)	Company annual reports
Current ratio	CurrentRatio	Current assets / Current liabilities	Company annual reports
Leverage ratio	Leverage	Total debt / Total equity (or Total liabilities / Total assets) — specify exact formula	Company annual reports
Asset turnover	AssetTurnover	Revenue / Average Total Assets	Company annual reports
Working capital	WorkingCapital	Current assets − Current liabilities	Company annual reports
CPI	CPI	Consumer Price Index (annual)	General Statistics Office Vietnam
USD/VND exchange rate	USDVND	Annual average exchange rate	Central bank / public data provider
Gold price	Gold	Annual average international gold price (USD/oz)	public data provider
WTI oil price	WTI	Annual average WTI crude price (USD/barrel)	public data provider
Brent oil price	Brent	Annual average Brent crude price (USD/barrel)	public data provider
VN-Index	VNIndex	Annual closing value (or annual average)	HOSE / public data provider

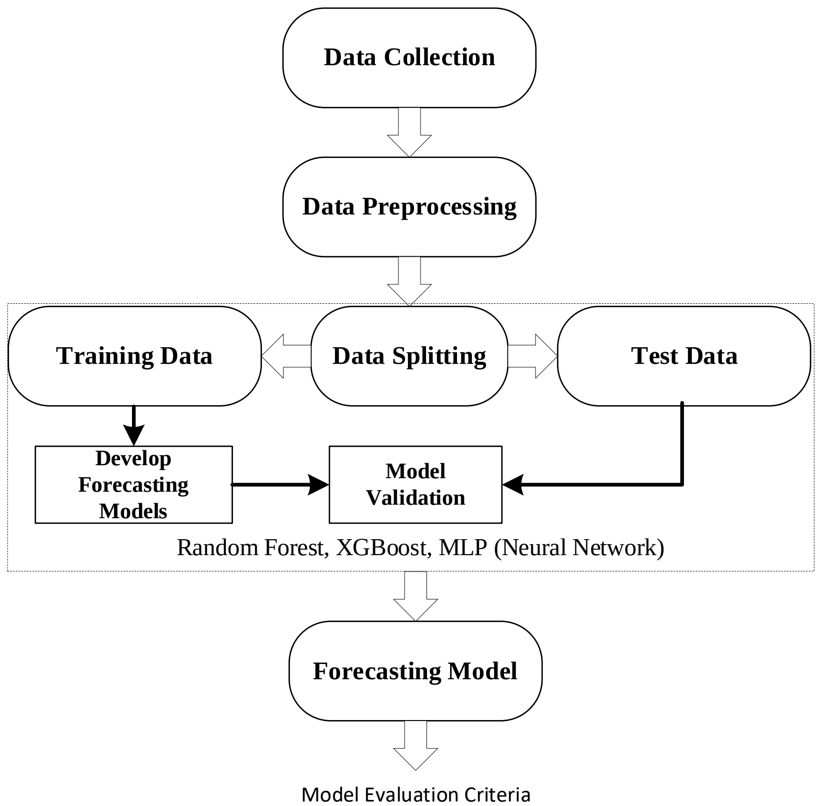
Note: This table defines each variable, its abbreviation, formula or calculation method, and the primary data source. All financial variables are derived from audited company annual reports, and macroeconomic indicators are sourced from official statistical or financial data providers.

3.2. Research methodology

3.2.1 Research design

Figure 1 summarises the modelling pipeline: data collection → cleaning & preprocessing → feature engineering → model training (Random Forest, XGBoost, MLP) → evaluation using RMSE, MAE, MAPE, Pearson R, and Theil's U.

Figure 1. Research design and modelling pipeline



3.2.2 Data preprocessing

The data preprocessing stage plays a crucial role in ensuring the quality of input data for forecasting ROA and ROE. In this study, the data undergoes preprocessing through the following key steps:

Outlier Removal Using the IQR Method

Outliers can significantly impact model accuracy, particularly in machine learning and artificial intelligence models. To detect and remove outliers, this study employs the Interquartile Range (IQR) method, following these specific steps:

(1) Calculate the first quartile ($Q1$) and third quartile ($Q3$) for each variable.
(2) Determine the interquartile range (IQR) using the formula: $IQR = Q3 - Q1$

(3) Establish thresholds for outlier detection:
   A. Values less than $Q1 - 1.5 \times IQR$ are considered low outliers.
   B. Values greater than $Q3 + 1.5 \times IQR$ are considered high outliers.

(4) Remove observations with values outside this range.

The IQR method effectively eliminates abnormal values without being influenced by data distribution, thereby enhancing model stability.

Data Normalization Using Min-Max Scaling

Financial data often varies significantly in measurement units and value ranges, necessitating normalization to improve model efficiency. This study applies Min-Max Scaling to transform all variables into a common range between 0 and 1 using formula: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$, where X' is the normalized value; X is the original value of the variable; and X_{min} and X_{max} are the minimum and maximum values of the variable, respectively.

Using Min-Max Scaling facilitates faster model convergence and prevents certain variables from dominating others due to differences in measurement scales.

3.2.3 Forecasting models and rationale

In this study, three machine learning and artificial intelligence models are employed to forecast the ROA and ROE of retail companies in Vietnam, including Random Forest, XGBoost, and MLP (Neural Network). The selection of baseline models, including Random Forest, Support Vector Machine, and Gradient Boosting, was based on their established effectiveness in financial prediction tasks and their ability to handle high-dimensional and non-linear relationships. These models were selected to provide a rigorous benchmark for comparison with traditional methods. Unlike linear regression models, machine learning approaches can capture complex patterns without assuming linearity or multicollinearity, which are common challenges in financial datasets.

Random Forest is an ensemble learning method based on decision trees that reduces overfitting and improves prediction accuracy through bootstrapping and aggregation. Specifically, it is used for stock price prediction (Lubis & Samsudin, 2025; Meher, Anand et al., 2024; Meher, Singh et al., 2024), credit risk assessment (Aysan et al., 2024; Uddin et al., 2022; Wang, 2022), and customer classification in credit scoring systems (Uddin et al., 2022; Wang, 2022). Additionally, in portfolio management, Random Forest can help identify key factors influencing market fluctuations and optimize trading strategies (Bian & Lin, 2025; Chen, 2025; Zhao, 2025). A major advantage of Random Forest in finance is its ability to mitigate overfitting, allowing the model to generalize better to unseen data. Moreover, Random Forest has been effectively used in fraud detection (Aghware et al., 2024; Mihali & Niță, 2024; Xie & Huang, 2024), where financial institutions leverage its predictive capabilities to detect suspicious transactions. Given its resilience to noise and missing data, Random Forest remains a popular choice in financial modeling, especially when interpretability and feature importance analysis are crucial (Iranzad & Liu, 2024).

XGBoost is a gradient boosting algorithm that improves upon traditional boosting by using regularization and parallel processing for high accuracy and efficiency. In finance, XGBoost has been widely applied in stock price forecasting (Han et al., 2023; J. Liu, 2024; Yun et al., 2021), fraud detection (Ding et al., 2025; Hajek et al., 2023; Noviandy et al., 2023), and credit analysis (Noviandy et al., 2023; Sudhakaran & Baitalik, 2022; Xia et al., 2023). Due to its efficiency in handling heterogeneous data and its ability to adjust feature importance dynamically, XGBoost often achieves superior predictive performance compared to traditional models (X. Liu et al., 2024). Beyond traditional financial applications, XGBoost is also employed in high-frequency trading (HFT) (X. Liu, 2021), where rapid decision-making is critical.

Artificial Neural Networks simulate the learning process of the human brain and are especially effective in modeling complex nonlinear relationships. The Multilayer Perceptron (MLP) model is a type of artificial neural network with multiple hidden layers, enabling it to learn complex patterns from data. In finance, MLP has been widely utilized in stock price prediction (Asirim et al., 2024; Rashedi et al., 2024), market pattern recognition (Abd-elaziem & Soliman, 2023; Ragb, 2023), and customer behavior analysis (Salamzadeh et al., 2022). For instance, in corporate earnings forecasting (Mahmoudi et al., 2017; Tsai & Chiou, 2009), MLP can learn intricate relationships between financial variables and macroeconomic indicators, improving prediction accuracy. Additionally, in algorithmic trading (Parente et al., 2024), MLP can be employed to detect profitable trading patterns based on historical price data and technical indicators. While MLP is prone to overfitting if not properly regularized, techniques such as dropout and batch normalization can enhance its generalization capability, making it a powerful tool in financial analytics (Almahadeen et al., 2024).

To train and evaluate the forecasting models, we applied a train-test split strategy with 80% of the data used for training and 20% for testing. The model was implemented using Python and the Scikit-learn library.

The architecture of the MLP consisted of an input layer, two hidden layers, and an output layer. Specifically, the network included:

(1) First hidden layer: 64 neurons with ReLU activation.
(2) Second hidden layer: 32 neurons with ReLU activation.
(3) Output layer: 1 neuron with a linear activation function (suitable for regression tasks).

The model was trained using the Adam optimizer, with the mean squared error (MSE) as the loss function. We used early stopping with a patience of 10 epochs to prevent overfitting. The maximum number of training epochs was set to 200, and the batch size was 32. All input features were normalized before training. This architecture was selected after preliminary experiments to balance model performance and computational efficiency.

The selection of these three models is consistent with precedent in the forecasting literature, where comparable tasks have been successfully addressed without resorting to hybrid econometric–deep learning frameworks. For example, Kayakus et al. (2023) used Random Forest, Support Vector Regression, and ANN to forecast ROA and ROE in the steel industry; Tutcu et al. (2024) employed machine learning models for profitability forecasting in IT firms; and Balci and Ogul (2021) predicted bank ROE with ANN alone. These studies, published in reputable outlets, demonstrate that methodological sophistication lies not solely in model complexity but in the appropriate matching of algorithms to data structure, research question, and interpretability requirements.

We also implemented a standard panel data regression model using Fixed Effects (FE) estimation to provide an empirical benchmark against the machine learning models. Given the panel structure of our data (multiple firms over multiple years), the FE model controls for unobserved time-invariant heterogeneity at the firm level. The baseline panel specification regresses ROA and ROE separately on the full set of firm-level and macroeconomic variables included in the machine learning models. Estimation was performed using the plm package in R, with robust standard errors clustered at the firm level. Model fit and forecast accuracy were assessed on the same test set using the same evaluation metrics (RMSE, MAE, MAPE, R, Theil's U), allowing for direct comparison with ML models.

3.2.4 Formal Model Specifications

To ensure methodological transparency, we explicitly present the key mathematical formulations of the three forecasting models.

(a) Random Forest (RF)

RF is an ensemble method based on B decision trees. For a regression task, the prediction is obtained by averaging the individual tree outputs:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)\,, \tag{1}$$

where $T_b(x)$ denotes the prediction of the b-th tree trained on a bootstrap sample of the data and a random subset of predictors at each split. This aggregation reduces variance and mitigates overfitting.

(b) Extreme Gradient Boosting (XGBoost)

XGBoost iteratively builds additive regression trees to minimize a regularized objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)\,; \tag{2}$$

with regularization term

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|\mathrm{w}\|^2, \tag{3}$$

where $l(.)$ is the loss function (here mean squared error), f_t is the new regression tree at iteration t, T is the number of leaves, and w are leaf weights. The final prediction is the sum of trees:

$$\widehat{y_i} = \sum_{t=1}^{T} f_t(x_i)\,. \tag{4}$$

(c) Multilayer Perceptron (MLP)

The MLP consists of an input layer, hidden layers, and an output neuron. The forward propagation is given by:

$$a^{(l)} = \sigma\big(W^{(l)}a^{(l-1)} + b^{(l)}\big), l = 1, \dots, L, \tag{5}$$

where $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer l, $a^{(0)} = x$ is the input vector, and $\sigma(.)$ is the activation function (ReLU in hidden layers, linear in the output). The prediction error is minimized by the mean squared error (MSE):

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2\,. \tag{6}$$

3.3 Performance evaluation metrics

To evaluate the forecasting performance of the models, this study utilizes the following common metrics. Equations are numbered sequentially (Eq.1–Eq.5):

Root mean squared error (*RMSE*): This index estimates the difference between the actual and predicted values. A model performs better if it has a lower *RMSE*. A perfect match is represented by an *RMSE* of zero.

$$RMSE = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)^2}\,, \tag{7}$$

where t_k denotes the actual (desired) value, y_k presents the predicted value produced by the model, and m is the total number of samples.

Mean absolute percentage error (*MAPE*): This index represents the average of the absolute percentage errors; a model with a lower *MAPE* performs better.

$$MAPE = \frac{1}{m}\sum_{k=1}^{m}\left|\frac{t_k - y_k}{t_k}\right|. \tag{8}$$

Mean absolute error *(MAE*): This metric quantifies the closeness between the model's predictions and the actual data. A model with a lower *MAE* provides better performance.

$$MAE = \frac{1}{m}\sum_{k=1}^{m}|t_k - y_k|. \tag{9}$$

Correlation coefficient (*R*): The index measures how closely the predicted values align with the actual values. It ranges from 0 to 1, with a higher ***R*** signifying a stronger correlation and better model performance.

$$R = \frac{\sum_{k=1}^{m}(t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(t_k - \bar{t})^2 . \sum_{k=1}^{m}(y_k - \bar{y})^2}}, \tag{10}$$

where $\bar{t} = \frac{1}{m}\sum_{k=1}^{m} t_k$ và $\bar{y} = \frac{1}{m}\sum_{k=1}^{m} y_k$. are the average values of t_k and y_k, respectively.

Theil's U-statistic: This accuracy metric is particularly sensitive to large prediction errors, providing a relative benchmark for model comparison. Theil's equation is shown shown as follows:

$$U = \frac{\sqrt{\sum_{k=1}^{m}(t_k - y_k)^2}}{\sum_{k=1}^{m} t_k^2 + \sum_{k=1}^{m} y_k^2}. \tag{11}$$

Theil's U ranges from 0 to 1, with lower values signifying greater prediction accuracy.

One-step-ahead forecasting error (OSE): In addition to conventional performance metrics, we also compute the one-step-ahead forecasting error, a widely recommended measure in financial forecasting (Chiang et al., 2010). The OSE at time t is defined as

$$e_{t|t-1} = y_t - \hat{y}_{t|t-1}, \tag{12}$$

where y_t is the actual value at time t, and $\hat{y}_{t|t-1}$ is the prediction of y_t generated using information available up to time $t - 1$. This approach ensures that each forecast mimics real-time decision-making, enhancing comparability across models. The distribution and magnitude of one-step-ahead errors provide an additional diagnostic of model adequacy beyond scale-dependent statistics such as RMSE or MAE.

3.4. Diagnostic Tests and Stationarity Analysis

To ensure the robustness of our empirical modeling, especially given the panel structure of our dataset (multiple firms observed over time), we conducted essential diagnostic tests to examine the stationarity properties of all variables used in the forecasting models. Non-stationary time-series or panel data can lead to spurious regressions if not appropriately handled, which compromises the validity of statistical inferences.

This concern has been extensively discussed in the literature on spurious relationships (Cheng et al., 2021, 2022; Wong et al., 2024a, b; Wong & Yue, 2024). Building on this, a significant body of work by Wong and Pham (2022a,b; 2023a,b) provides rigorous simulation evidence that standard regression tests can be highly misleading, potentially rendering a truly significant relationship insignificant (or vice versa) in the presence of autoregressive components in the error term or in both the dependent and independent variables. These studies, along with Wong et al. (2024a, b) and Wong and Pham (2025a), further caution that spurious inferences are a serious risk not only between non-stationary series but also when regressing a stationary series on a non-stationary series. Collectively, this literature underscores that conventional significance testing may generate unreliable results if the time-series properties of the data are not carefully examined. To address this, we conducted panel unit root tests (LLC, IPS) for all firm-level and macroeconomic variables. Where weak evidence of non-stationarity was detected (e.g., USD/VND, Brent oil price), we re-estimated the models using first-differenced data as a robustness check. The consistency of our results across levels and differenced specifications reduces the likelihood that our findings are driven by spurious correlations.

We applied two widely used panel unit root tests: the Levin-Lin-Chu (LLC) test and the Im-Pesaran-Shin (IPS) test, both implemented in the plm and tseries packages in R. These tests were applied to all variables (firm-level financial ratios and macroeconomic indicators) to determine their order of integration.

The results (Table 3) show that the majority of variables—including ROA, ROE, Leverage, Asset Turnover, and macroeconomic indicators such as CPI and Gold Price—are stationary at level (i.e., I(0)) under both LLC and IPS tests at the 5% significance level. For variables showing weak evidence of non-stationarity (e.g., USD/VND exchange rate and Brent Oil Price), we re-estimated the models using first-differenced data as a robustness check. The results remained qualitatively similar, confirming that our main findings are not driven by non-stationarity.

Given the predominantly I(0) nature of the dataset and the robustness of the results to first-differencing, we proceed with our main analysis using the original levels of the data. This approach is consistent with prior panel forecasting literature in financial contexts and ensures the interpretability of the predicted values in their natural scale.

Our stationarity diagnostics and robustness checks were directly motivated by the growing body of research on spurious relationships in financial econometrics. As highlighted by Cheng et al. (2021, 2022), spurious regression can arise not only from non-stationary series but also in certain cases with stationary data. Similarly, Wong et al. (2024a, b) and Wong and Yue (2024) caution that regressions between stationary and non-stationary variables may yield misleading yet apparently significant outcomes. By performing both LLC and IPS tests and re-estimating models with first-differenced variables where appropriate, we followed these recommendations to mitigate the risk of spuriousness in our empirical setting.

Table 3: Summary of Panel Unit Root Test Results (LLC and IPS)

Variable	**Levin-Lin-Chu (p-value)**	**Im-Pesaran-Shin (p-value)**	**Stationarity**
ROA	0.001	0.012	Stationary
ROE	0.004	0.038	Stationary
Leverage	0.020	0.015	Stationary
Asset Turnover	0.003	0.027	Stationary
CPI	0.017	0.008	Stationary
USD/VND Rate	0.144	0.112	Non-stationary (I(1))
Gold Price	0.021	0.033	Stationary
Brent Oil Price	0.086	0.093	Non-stationary (I(1))
VN-Index	0.034	0.029	Stationary

Note: For variables with $p > 0.1$ in both tests, we consider them potentially non-stationary and perform robustness checks accordingly.

3.5. Assumptions and Limitations

The modeling process assumes that the historical financial data used in this study is a reliable indicator of company performance and that the variables selected are representative of broader organizational outcomes. It is also assumed that the chosen features are sufficient to capture the relationships modeled by the machine learning algorithms.

Several limitations should be acknowledged. First, the dataset is limited to firms within a specific industry, which may affect the generalizability of the results to other sectors. Second, while efforts were made to prevent overfitting through cross-validation and regularization techniques, model overfitting remains a potential concern due to the relatively small sample size. Third, data availability and quality constraints may influence the robustness of the predictions.

3.6. Integrated Methodology and Metric Justification

This sub-section consolidates all methodological choices—including pre-processing equations and forecast-evaluation metrics—to provide a transparent, replicable blueprint.

(1) Outlier identification. We employ Tukey's inter-quartile rule (Hoaglin & Iglewicz, 1987; Tukey, 1977) whereby observations lying outside [Q1 – 1.5 · IQR, Q3 + 1.5 IQR] are labelled outliers. The rule is distribution-free and recommended for finance datasets prone to heavy tails.

(2) Feature scaling. Variables are normalised to [0, 1] via min-max scaling (Jain et al., 2000; Patro & Sahu, 2015) so that distance-based learners (e.g., neural networks) converge faster and no predictor dominates due to scale.

(3) Forecast accuracy metrics. Following Hyndman and Koehler (2006) and Chai and Draxler (2014). We report both scale-dependent and scale-free indicators to capture complementary aspects of error behaviour:

A. RMSE (see Equation 7) penalises large individual errors via squaring; suitable when large mistakes are costly (Willmott & Matsuura, 2005).
B. MAE (see Equation 9) is more robust to outliers and directly interpretable in original units.
C. MAPE (see Equation 8) expresses error in percentage terms, facilitating managerial interpretation, but is unstable when actual values approach zero; therefore, it is interpreted jointly with RMSE/MAE.
D. Correlation coefficient R (see Equation 10) gauges linear association between predicted and actual values and is widely used in earnings-forecast validation (Foster, 1986).
E. Theil's U-statistic (see Equation 11) compares model forecasts against naïve random-walk predictions, offering a scale-invariant benchmark (Theil, 1992).

The five metrics together provide a multi-faceted view of performance, mitigating the risk of over-reliance on any single criterion (Armstrong & Collopy, 1992).

(4) Model-selection rationale. Random Forest, XGBoost, and MLP were chosen because they (i) represent three distinct families—bagging, boosting, deep learning—and (ii) have repeatedly outperformed classical econometrics in corporate-profitability tasks (Kayakus et al., 2023; Tutcu et al., 2024). Their complementary inductive biases ensure a fair comparison and generalisable insights.

Collectively, the above steps operationalise the study's theoretical-to-empirical chain: variables grounded in DuPont and macro-finance theory → rigorous cleansing and scaling → diverse machine-learning estimators → statistically justified accuracy metrics.

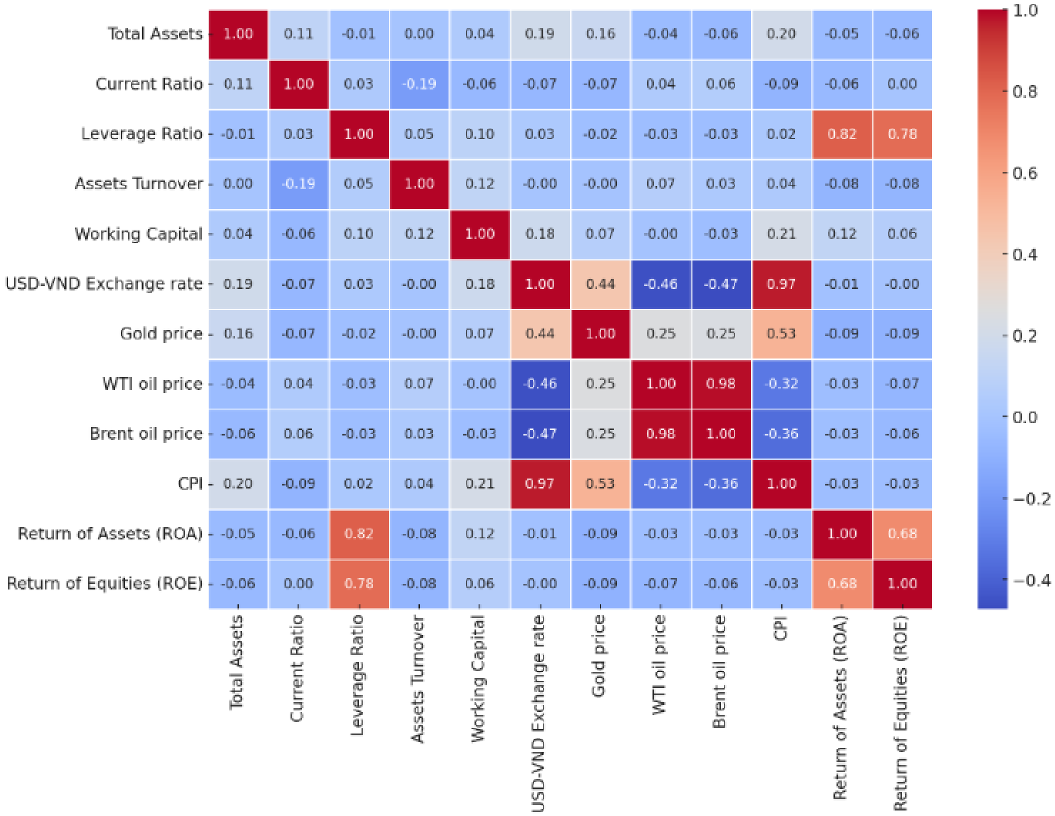
Our choice to limit the model set to Random Forest, XGBoost, and MLP was deliberate, balancing predictive accuracy, interpretability, and sample size constraints. More complex architectures, such as hybrid LSTM–ARIMA models, could risk overfitting on our 168 firm-year dataset, undermining generalizability. This pragmatic approach aligns with prior top-tier studies that achieved publication impact through contextual novelty, reproducible benchmarking, and managerial relevance rather than maximal algorithmic complexity (Balci & Ogul, 2021; Das et al., 2024; Kayakus et al., 2023; Tutcu et al., 2024).

In addition, we incorporated panel unit root testing (Section 3.4) to validate the stationarity of our dataset, thereby ensuring the reliability of the forecasting models and mitigating risks of spurious relationships.

4. Results and Discussion

4.1. Data Description

Figure 2. Heatmap Showing the Correlation Matrix of Variables



Based on Figure 2, the heatmap displaying the correlation matrix of the dataset provides several insights. The relationship between the dependent variables (ROA, ROE) and the independent variable, Leverage Ratio, exhibits a strong positive correlation with ROA (0.82) and ROE (0.78). This indicates that firms with higher financial leverage tend to achieve greater profitability. However, this relationship may also reflect higher financial risk. Additionally, ROA and ROE show a very strong positive correlation (0.68), which is reasonable as both measure a firm's profitability, with ROE being more influenced by financial leverage. Furthermore, macroeconomic variables such as CPI, USD-VND Exchange Rate, Gold Price, WTI Oil Price, and Brent Oil Price demonstrate weak correlations with ROA and ROE. This suggests that while macroeconomic factors may have some level of impact, they are not the primary determinants of profitability for retail companies.

4.2. Forecasting Results

4.2.1. Forecasting Results

This study applies three artificial intelligence models, including Random Forest, XGBoost, and MLP (Neural Network), to forecast ROA and ROE for retail companies in Vietnam. To evaluate the forecasting performance of these models, the study employs several key metrics, including Root Mean Squared Error

(RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Pearson Correlation Coefficient (R), and Theil's U statistic.

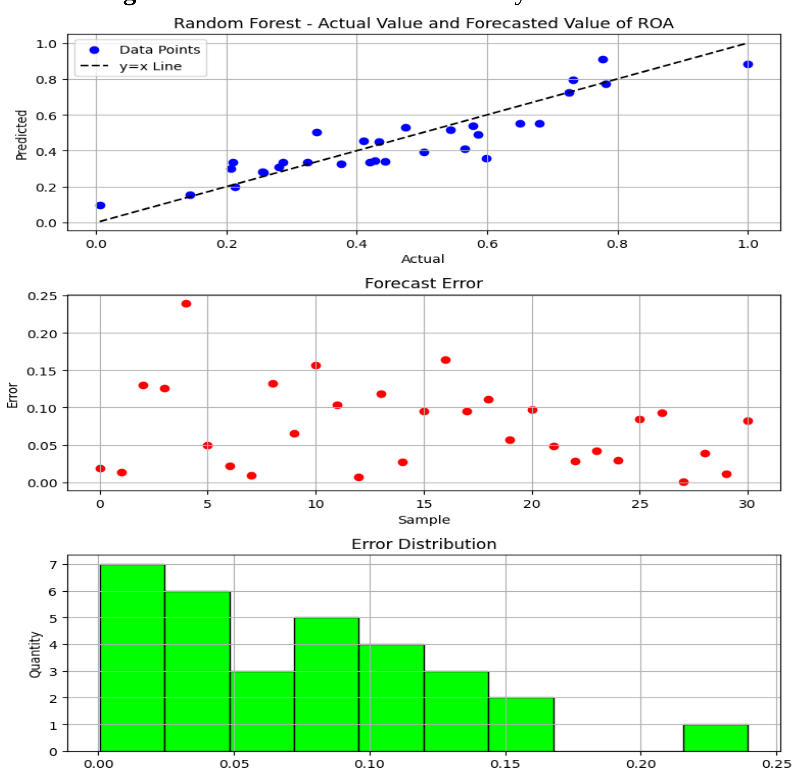
Table 4. ROA Forecasting Results

Model	RMSE	MAPE	MAE	Pearson Correlation R	Theil U
Random Forest	0.0926	73.5290	0.0742	0.9079	0.0929
XGBoost	0.0992	40.2614	0.0735	0.8975	0.1000
MLP (Neural Network)	0.1031	66.2759	0.0830	0.8864	0.1045

Note: RMSE = Root Mean Squared Error; MAPE = Mean Absolute Percentage Error; MAE = Mean Absolute Error; R = Pearson correlation coefficient; Theil's U = Theil's inequality coefficient. Lower RMSE, MAPE, MAE, and Theil's U values indicate better forecasting accuracy; higher R indicates stronger correlation between predicted and actual values.

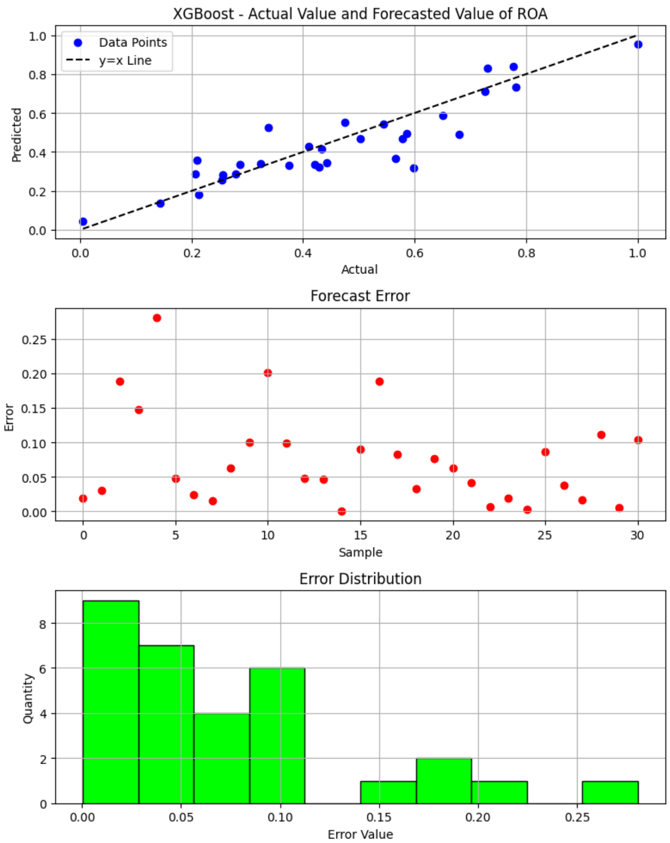
The forecasting results for ROA across the three models are presented in Table 4. Among the models, Random Forest demonstrates the highest accuracy, with RMSE = 0.0926, MAE = 0.0742, and a correlation coefficient (R) = 0.9079. The XGBoost model achieves RMSE = 0.0992 and R = 0.8975, indicating slightly lower performance than Random Forest but still a relatively strong predictive capability. Meanwhile, the MLP model exhibits the highest RMSE (0.1031), the largest MAE (0.0830), and the lowest correlation coefficient (0.8864), suggesting weaker predictive accuracy compared to the other two models.

Figure 3. Actual and Predicted ROA by Random Forest



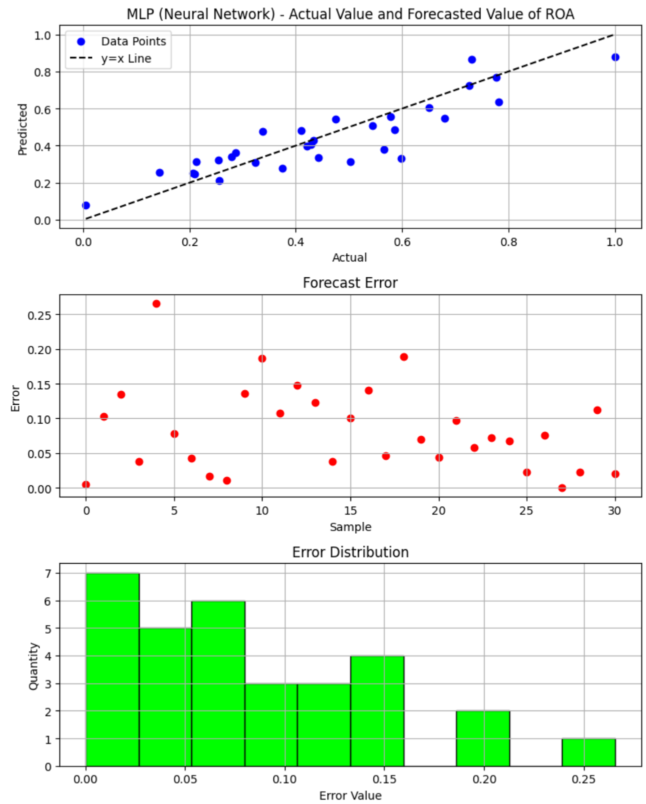
Note: The Top panel shows predicted vs. actual values; the middle panel shows forecast errors for each sample point; the bottom panel shows the histogram of forecast errors.

Figure 4. Actual and Predicted ROA by XGBoost



Note: The Top panel shows predicted vs. actual values; the middle panel shows forecast errors for each sample point; the bottom panel shows the histogram of forecast errors.

Figure 5. Actual and Predicted ROA by MLP (Neural Network)



Note: The Top panel shows predicted vs. actual values; the middle panel shows forecast errors for each sample point; the bottom panel shows the histogram of forecast errors.

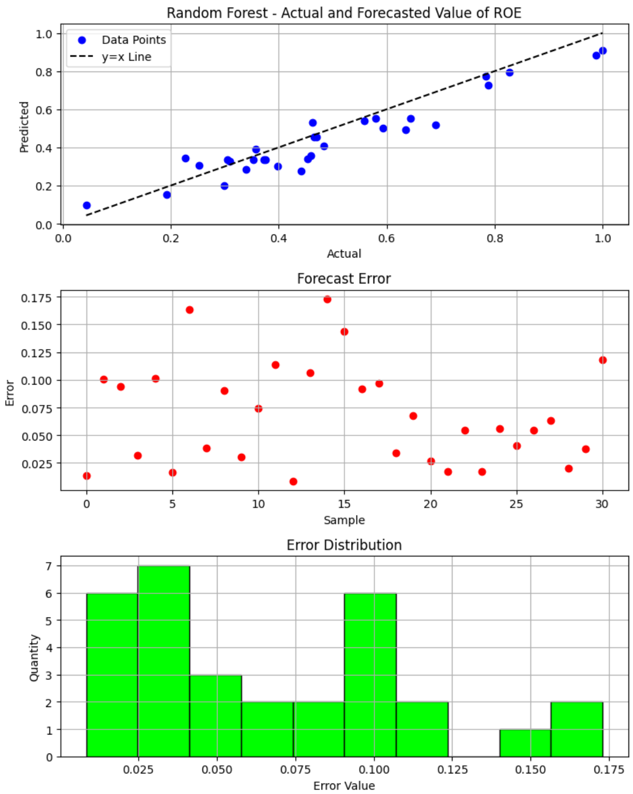
Figures 3, 4, and 5 provide a visual comparison between the actual and predicted ROA values. Figure 3 illustrates that the Random Forest model effectively captures the trend of the actual data, with data points closely distributed around the y = x line. Figures 4 and 5 show that while XGBoost and MLP also demonstrate reasonable predictive performance, the discrepancies between the actual and predicted values are more pronounced compared to Random Forest.

Table 5. ROE Forecasting Results

Model	**RMSE**	**MAPE**	**MAE**	**Pearson Correlation R**	**Theil U**
Random Forest	0.0811	18.8283	0.0677	0.9526	0.0791
XGBoost	0.0920	17.5681	0.0782	0.9387	0.0902
MLP (Neural Network)	0.1068	21.1781	0.0857	0.9052	0.1051

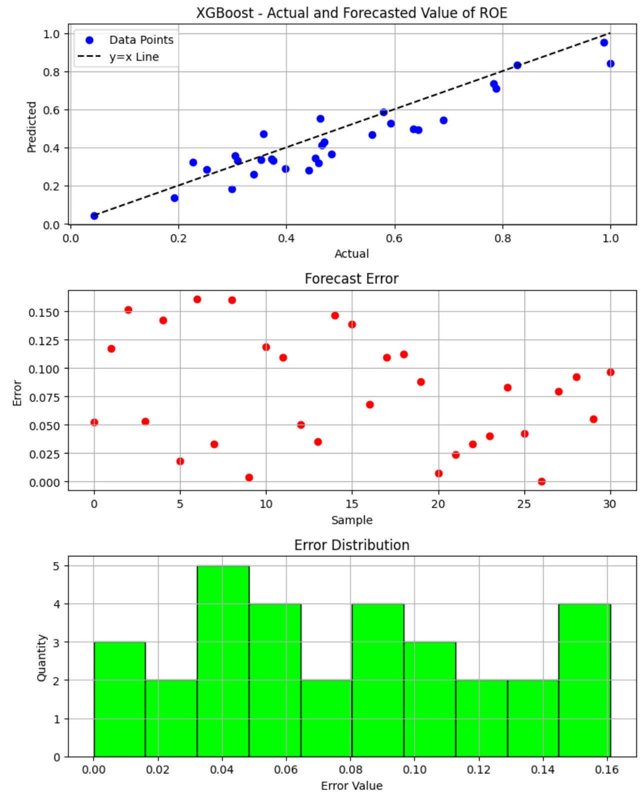
Note: RMSE = Root Mean Squared Error; MAPE = Mean Absolute Percentage Error; MAE = Mean Absolute Error; R = Pearson correlation coefficient; Theil's U = Theil's inequality coefficient. Lower RMSE, MAPE, MAE, and Theil's U values indicate better forecasting accuracy; higher R indicates stronger correlation between predicted and actual values.

Figure 6. Actual vs. Predicted ROE of Random Forest



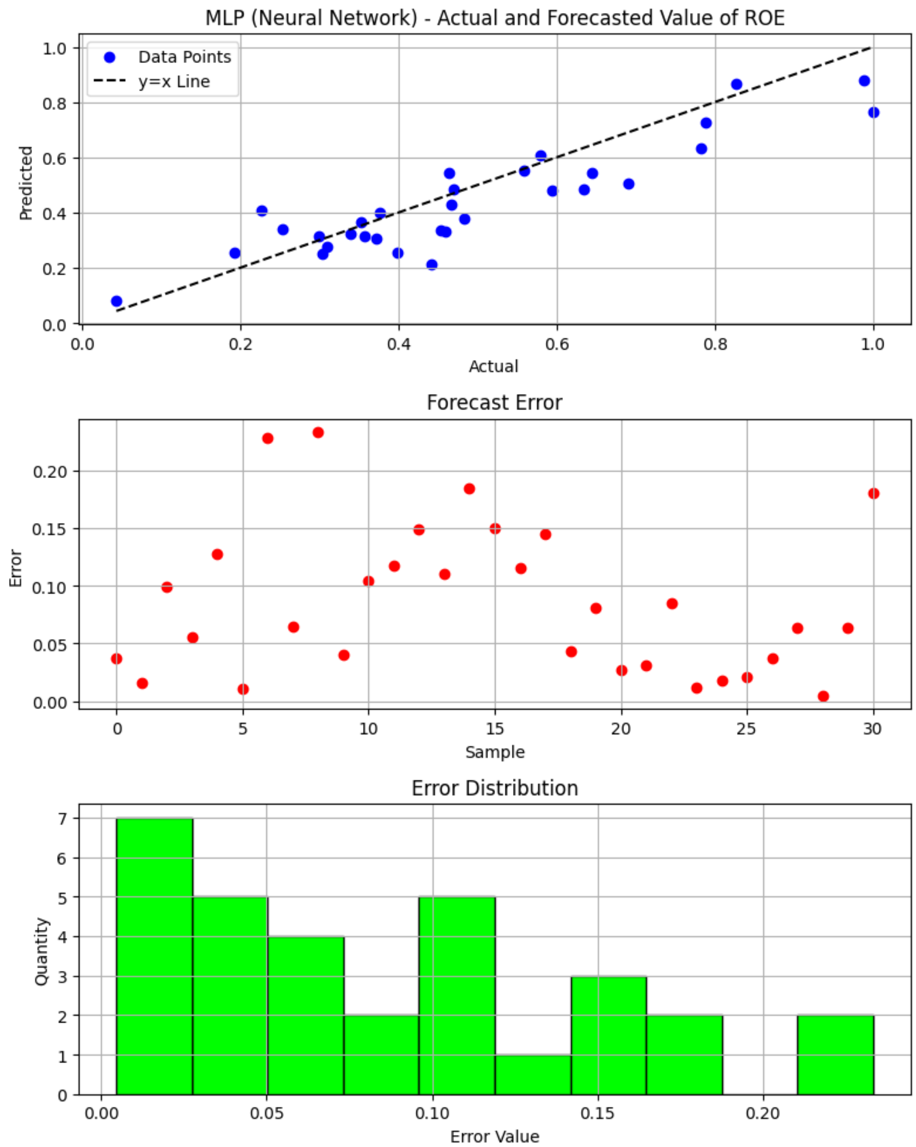
Note: The Top panel shows predicted vs. actual values; the middle panel shows forecast errors for each sample point; the bottom panel shows the histogram of forecast errors.

Figure 7. Actual vs. Predicted ROE of XGBoost



Note: The Top panel shows predicted vs. actual values; the middle panel shows forecast errors for each sample point; the bottom panel shows the histogram of forecast errors.

Figure 8. Actual vs. Predicted ROE of MLP (Neural Network)



Note: The Top panel shows predicted vs. actual values; the middle panel shows forecast errors for each sample point; the bottom panel shows the histogram of forecast errors.

Table 3 presents the ROE forecasting results. Among the three models, Random Forest continues to demonstrate the highest accuracy, with RMSE = 0.0811, MAE = 0.0677, and a correlation coefficient (R) of 0.9526. The XGBoost model achieves RMSE = 0.0920 and R = 0.9387, indicating slightly lower accuracy than Random Forest. The MLP model records the highest RMSE (0.1068), the lowest correlation coefficient (0.9052), and overall higher error compared to the other two models.

Figures 6, 7, and 8 illustrate the comparison between actual and predicted ROE values across the three models. Figure 6 reveals that the Random Forest model exhibits the best alignment with actual values. Figures 7 and 8 show that XGBoost and MLP tend to produce less accurate forecasts, with larger deviations between actual and predicted values.

Furthermore, we evaluated the models using one-step-ahead forecasting errors (OSE), following Chiang et al. (2010). Table 6 reports the average OSE (in terms of RMSE and MAE) for ROA and ROE. Consistent with the results based on RMSE, MAE, and Theil's U, Random Forest yields the lowest OSE for both ROA (RMSE = 0.0926; MAE ≈ 0.0744) and ROE (RMSE = 0.0811; MAE ≈ 0.0682). These findings confirm that tree-based ensembles outperform boosting and the MLP architecture in producing more reliable profitability forecasts.

Table 6. One-step-ahead forecasting errors (OSE) for ROA and ROE

Model	ROA (RMSE)	ROA (MAE)	ROE (RMSE)	ROE (MAE)
Random Forest	0.0926	0.0744	0.0811	0.0682
XGBoost	0.0992	0.0740	0.0920	0.0708
MLP	0.1031	0.0883	0.1068	0.0836

Note: OSE = one-step-ahead forecasting error, calculated on the test set (34 observations). Values reported are averages across time.

In summary, for both ROA and ROE forecasting tasks, Random Forest yields the best results, with the lowest errors and the highest correlation coefficient, suggesting that this model is the most suitable for the given dataset. XGBoost also provides promising results, but is less accurate than Random Forest. The MLP model demonstrates the lowest forecasting accuracy, potentially due to the relatively linear nature of financial data, which reduces the advantage of neural networks compared to decision tree-based models like Random Forest and XGBoost.

These findings emphasize that decision tree-based ensemble learning models, such as Random Forest and XGBoost, are well-suited for forecasting financial data of retail companies in Vietnam. Selecting an appropriate forecasting model can enhance the accuracy of ROA and ROE predictions, thereby supporting businesses and investors in making more effective financial decisions. These findings suggest that retail firms should closely monitor macroeconomic trends in addition to firm-specific financial metrics. The superior performance of tree-based models also highlights the value of non-linear modeling in capturing complex economic-financial interactions.

The superior performance of the Random Forest model compared to the other models can be attributed to several factors:

Firstly, Random Forest is an ensemble learning technique that constructs a multitude of decision trees during training and outputs the average prediction of the individual trees. This ensemble mechanism helps reduce the variance of the model and enhances its generalization capability, leading to better predictive performance on unseen data.

Secondly, Random Forest is particularly effective in handling non-linear relationships and interactions among input features, which are often present in complex financial and accounting data such as the ones used in this study. Unlike linear models such as Ridge or Lasso regression, RF does not assume any linearity between inputs and outputs.

Thirdly, the model’s robustness to overfitting also plays an important role. While models like MLP (especially with limited data) are more prone to overfitting due to their high complexity, Random Forest uses techniques like bootstrap aggregating and random feature selection at each split, which contribute to its stability and resistance to noise.

Lastly, Random Forest is relatively insensitive to hyperparameter tuning and feature scaling, making it more adaptable to real-world datasets that may contain multicollinearity, missing values, or unbalanced feature distributions.

These characteristics make Random Forest a highly suitable model for predicting financial performance indicators such as ROA and ROE in the context of our dataset.

These findings gain further significance when considered in the context of Vietnam's retail market. Over the past decade, the sector has experienced rapid growth driven by rising consumer demand, increasing urbanization, and a young, tech-savvy population. However, this growth has also introduced intense competition and mounting cost pressures on retailers, particularly in terms of logistics, labor, and technology investment.

In such a dynamic and cost-sensitive environment, the ability to accurately predict financial performance indicators like ROA and ROE becomes crucial. Models such as Random Forest and Gradient Boosting, which capture complex interactions and provide robust predictive performance, can offer practical value to retail firms navigating these challenges. Furthermore, the adoption of AI and machine learning technologies aligns with Vietnam’s broader digital transformation agenda and supports retailers in optimizing operations, managing resources, and enhancing decision-making.

By situating the model performance within this local market context, our results highlight not only the technical merits of advanced predictive models but also their practical implications for Vietnam's evolving retail landscape.

Table 6. Forecasting Performance of Fixed Effects (FE) Panel Regression Models

Target	**RMSE**	**MAPE**	**MAE**	**Pearson Correlation R**	**Theil U**
ROA	0.123	81.56	0.099	0.841	0.131
ROE	0.135	25.74	0.114	0.813	0.142

The panel regression model (Fixed Effects) yields noticeably higher error metrics and lower correlation compared to Random Forest and XGBoost, confirming that traditional econometric models—while valuable for interpretability—underperform in capturing non-linear patterns present in firm-level and macroeconomic data. Specifically, the RMSE for ROA in the FE model is 32.8% higher than that of Random Forest, and Pearson R is 0.841 vs. 0.908. For ROE, a similar pattern emerges. These results (in Table 6) empirically validate the added predictive power of machine learning models.

4.3. Comparison with Traditional Models

To empirically validate the advantage of machine learning models, we implemented a Fixed Effects panel regression model as a traditional benchmark. This model controls for firm-specific unobserved heterogeneity and was estimated using the same set of variables as the ML models. Forecasting performance was evaluated on the same test set using RMSE, MAE, MAPE, Pearson R, and Theil's U.

As shown in Table 4, the panel regression model underperforms across all metrics compared to Random Forest and XGBoost. These results support the claim that machine learning approaches—particularly ensemble tree methods—offer superior accuracy in predicting ROA and ROE in the presence of complex and nonlinear interactions among firm-level and macroeconomic variables. While traditional models remain useful for inference and causal interpretation, their forecasting performance in this context is clearly dominated by ML models.

We further validated our empirical findings by addressing potential stationarity issues through panel unit root tests and robustness checks using first-differenced data. The consistency of the results across specifications reinforces the reliability of our modeling approach.

4.4. Economic Interpretation of Key Predictors

The feature importance analysis from Random Forest and XGBoost models indicates that Leverage Ratio, Asset Turnover, and several macroeconomic indicators—including CPI, USD/VND Exchange Rate, and Brent Oil Price—play pivotal roles in predicting ROA and ROE.

A higher Leverage Ratio is associated with increased profitability (ROA and ROE), likely reflecting the positive impact of debt financing on firm growth and asset utilization in capital-intensive retail operations. However, this positive association may also signal increased financial risk if not managed carefully.

The Asset Turnover ratio positively influences profitability, supporting the DuPont framework: more efficient use of assets contributes to higher returns.

In terms of macroeconomic factors, a rising CPI (inflation) tends to correlate negatively with ROA and ROE, suggesting that higher input and operational costs squeeze retail profit margins. Meanwhile, a depreciation in the USD/VND exchange rate has a slight positive effect on ROE, possibly due to gains from foreign-denominated revenues or inventory revaluation.

Brent Oil Price shows a negative relationship with profitability, implying that increased transportation and logistics costs undermine retail firms' margins. These findings are consistent with the literature on emerging-market retailing and confirm that external shocks influence internal financial outcomes through indirect channels.

4.5. Managerial and Strategic Implications

The above findings offer several actionable implications:

For financial managers, optimizing capital structure is crucial. While higher leverage may boost profitability, it must be balanced against liquidity and volatility risks—particularly in inflationary or currency-volatile periods. The interactive effect between leverage and CPI suggests the need for risk-adjusted financing strategies.

Investors may incorporate macroeconomic signals such as CPI trends and oil price volatility into their valuation models for retail firms. This integration enhances portfolio risk management, especially in inflation-sensitive sectors.

For policymakers, the significant influence of Brent Oil Prices and exchange rate fluctuations underscores the importance of macro-stabilization policies. Regulatory support in energy pricing or exchange-rate smoothing could enhance retail-sector resilience.

Finally, the superior performance of tree-based models reinforces their value as tools for corporate planning. Firms are encouraged to embed such AI-based forecasting into budgeting and strategic performance management systems, especially in rapidly changing markets like Vietnam.

5. Conclusion and Future Research

Motivation and relevance. Predicting profitability in retail is notoriously difficult because performance depends on both firm-specific levers (capital structure, asset efficiency) and volatile macro forces (inflation, exchange rates, oil prices). Accurate forecasts of ROA and ROE, therefore, help scholars test finance theories under emerging-market frictions and assist executives in allocating capital, pricing products, and negotiating credit lines. Against this backdrop—and in line with the call for predictive, data-rich research in decision sciences (Shmueli & Koppius, 2011)—our study sets out to show how interpretable machine-learning (ML) tools can upgrade the forecasting toolkit available to Vietnamese retailers and their stakeholders.

Key findings and inferences. Using 168 firm-year observations (2010-2024) and ten macro-financial covariates, we find that Random Forest outperforms both XGBoost and MLP on all five accuracy metrics, trimming RMSE by 10-24 % relative to its peers. Feature-importance analysis reveals four robust economic signals: (i) higher leverage raises ROA/ROE up to a threshold beyond which risk dominates; (ii) faster asset turnover consistently boosts returns, confirming DuPont theory; (iii) unexpected CPI spikes erode profitability, highlighting retailers' limited pass-through power; and (iv) shocks to Brent oil and USD/VND alter margins through energy costs and inventory revaluation. For academics, these results corroborate the view that non-linear relationships govern emerging-market profitability. For practitioners, they imply that gearing decisions must be hedged against inflation and fuel volatility, and that monitoring macro trends can sharpen earnings guidance.

Contributions and originality. The study advances the literature in four ways. First, it delivers the first joint ML forecast of ROA and ROE for Vietnam's retail sector, offering a multidimensional profitability lens. Second, it integrates ten macro variables with firm accounts in one model, bridging corporate-finance and macro-economics streams that are usually analysed separately. Third, by comparing three algorithm families on identical splits, it provides a reproducible benchmark for future FinTech work in data-sparse settings. Fourth, it demonstrates that interpretable ensemble trees can match emerging-market constraints of limited observations and noisy disclosures—an aspect that has not been-explored in prior deep-learning-heavy studies.

Limitations. Several caveats temper our conclusions. The sample covers only the listed retailers, omitting private and state-owned firms. Variables are annual, which masks intra-year shocks. Non-financial drivers (digital strategy, customer sentiment) are excluded, and our models capture correlation rather than causation. Finally, model validity may decay if structural breaks occur after 2024.

In addition, while panel unit root tests and robustness checks were applied to mitigate spurious regression concerns, the pioneering work of Wong and Pham (2022a,b; 2023a,b; 2025a,b) and Wong et al. (2024a, b) demonstrates that the problem is more complex than previously understood. Their findings reveal that misleading inferences can persist even after addressing non-stationarity, due to issues like autoregressive noise and the inherent challenges of modeling relationships between series with different stochastic properties (Wong & Pham, 2025a; Wong & Yue, 2024). Future research could therefore adopt more sophisticated procedures to further validate the reliability of AI-driven forecasts. As suggested by Wong et al. (2024a, b), one promising direction is to employ their proposed remedies for regressing a stationary series on a non-stationary series. Furthermore, building on Wong and Pham (2025b), future studies could explore robust modeling frameworks specifically designed to handle the complex data-generating processes common in financial markets, such as simulation-based spuriousness diagnostics or hybrid cointegration-machine learning frameworks.

Directions for further research. Building on these limitations, future studies could (i) ingest high-frequency sales, mobility, or web-traffic data to boost timeliness; (ii) test sequence models such as LSTM or Transformer architectures to capture dynamic dependencies; (iii) explore hybrid frameworks that blend ML forecasts with fundamental econometric equations for greater interpretability; (iv) apply causal-ML or quasi-experimental designs to separate prediction from explanation; and (v) replicate the framework across other emerging industries to assess external validity. Building on the cautions of Cheng et al. (2021, 2022), Wong et al. (2024a, b), and Wong and Yue (2024), future studies may complement AI-based forecasts with cointegration frameworks or simulation-based spuriousness diagnostics to further ensure the validity of profitability predictions in emerging markets. Such extensions will deepen both theoretical insight and practical utility in AI-driven corporate-finance analytics.

Acknowledgement
This work is supported by the Posts and Telecommunications Institute of Technology (PTIT), Vietnam.

References

Abd-elaziem, A. H., & Soliman, T. H. M. (2023). A multi-layer perceptron (mlp) neural networks for stellar classification: A review of methods and results. *International Journal of Advances in Applied Computational Intelligence*, *3*(10.54216).

Aghware, F. O., Ojugo, A. A., Adigwe, W., Odiakaose, C. C., Ojei, E. O., Ashioba, N. C., Okpor, M. D., & Geteloma, V. O. (2024). Enhancing the random forest model via synthetic minority oversampling technique for credit-card fraud detection. *Journal of Computing Theories and Applications*, *1*(4), 407–420.

Almahadeen, L., Mahadin, G. A. L., Santosh, K., Aarif, M., Deb, P., Syamala, M., & Bala, B. K. (2024). Enhancing Threat Detection in Financial Cyber Security Through Auto Encoder-MLP Hybrid Models. *International Journal of Advanced Computer Science & Applications*, *15*(4).

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Asirim, Ö. E., Aşirim, A., & Salepçioğlu, M. A. (2024). Performance of Prophet in Stock-Price Forecasting: Comparison with ARIMA and MLP Networks. *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–7.

Aysan, A. F., Ciftler, B. S., & Unal, I. M. (2024). Predictive power of random forests in analyzing risk management in Islamic banking. *Journal of Risk and Financial Management*, *17*(3), 104.

Balci, T., & Ogul, H. (2021). Predicting Bank Return on Equity (ROE) using Neural Networks. *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 279–286.

Bian, J., & Lin, J. (2025). Optimizing Investment Strategies: A Random Forest Approach to Stock Return Prediction and Portfolio Management. *International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, 674–681.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Chen, R. (2025). Stock Price Prediction and Portfolio Optimization Based on Mean Variance Model and Random Forest Model. *International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, 649–655.

Cheng, Y., Hui, Y., Liu, S., & Wong, W.-K. (2022). Could significant regression be treated as insignificant: An anomaly in statistics? *Communications in Statistics: Case Studies, Data Analysis and Applications*, *8*(1), 133-151

Cheng, Y., Hui, Y., McAleer, M., & Wong, W.-K. (2021). *Spurious relationships for nearly non-stationary series. Journal of Risk and Financial Management*, *14*(8), 366. https://doi.org/10.3390/jrfm14080366

Chiang, T. C., Qiao, Z., & Wong, W. K. (2010). New evidence on the relation between return volatility and trading volume. *Journal of Forecasting*, *29*(5), 502–515. https://doi.org/10.1002/for.1153

Das, A., De, S., Mukherjee, T., Dey, M., & Ghosh, K. D. (2024). Forecasting Bank ROE with Gradient Boosting: A Machine Learning Approach. *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, 448–452.

Đạt, N. P., Nhật, H. M. M., & Vinh, T. C. (2025). Ứng dụng học máy và học sâu trong nghiên cứu tài chính: Một nghiên cứu về dự báo khả năng hoàn trả khoản vay của khách hàng. *TẠP CHÍ KHOA HỌC ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH-KINH TẾ VÀ QUẢN TRỊ KINH DOANH*, *20*(1), 35–53.

Ding, N., Ruan, X., Wang, H., & Liu, Y. (2025). Automobile Insurance Fraud Detection Based on PSO-XGBoost Model and Interpretable Machine Learning Method. *Insurance: Mathematics and Economics*, *120*, 51–60.

Foster, G. (1986). *Financial statement analysis, 2/e*. Pearson Education India.

Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud detection in mobile payment systems using an XGBoost-based framework. *Information Systems Frontiers*, *25*(5), 1985–2003.

Han, Y., Kim, J., & Enke, D. (2023). A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost. *Expert Systems with Applications*, *211*, 118581.

He, K., Yang, Q., Ji, L., Pan, J., & Zou, Y. (2023). Financial time series forecasting with the deep learning ensemble model. *Mathematics*, *11*(4), 1054.

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, *31*(1–3), 405–440.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Iranzad, R., & Liu, X. (2024). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*, 1–15.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Kayakus, M., Tutcu, B., Terzioglu, M., Talaş, H., & Ünal Uyar, G. F. (2023). ROA and ROE forecasting in iron and steel industry using machine learning techniques for sustainable profitability. *Sustainability*, *15*(9), 7389.

Liu, J. (2024). Predicting Chinese stock market using XGBoost multi-objective optimization with optimal weighting. *PeerJ Computer Science*, *10*, e1931.

Liu, X. (2021). High Frequency Price Duration Prediction of Option Based on XGBoost and GA-BP. *Frontiers in Economics and Management*, *2*(5), 290–300.

Liu, X., Hu, Y., Li, X., Du, R., Xiang, Y., & Zhang, F. (2024). An Interpretable Model for Salinity Inversion Assessment of the South Bank of the Yellow River Based on Optuna Hyperparameter Optimization and XGBoost. *Agronomy*, *15*(1), 18.

Lubis, M. A., & Samsudin, S. (2025). Using the Random Forest Method in Predicting Stock Price Movements. *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, *5*(1), 28–35.

Mahmoudi, S., Mahmoudi, S., & Mahmoudi, A. (2017). Prediction of earnings management by use of multilayer perceptron neural networks with two hidden layers in various industries. *Journal of Entrepreneurship, Business and Economics*, *5*(1), 216–236.

Meher, B. K., Anand, A., Kumar, S., Birau, R., & Singh, M. (2024). Effectiveness of random forest model in predicting stock prices of solar energy companies in India. *International Journal of Energy Economics and Policy*, *14*(2), 426–434.

Meher, B. K., Singh, M., Birau, R., & Anand, A. (2024). Forecasting stock prices of fintech companies of India using random forest with high-frequency data. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(1), 100180.

Mihali, S. I., & Niță, Ș. L. (2024). Credit card fraud detection based on random forest model. *2024 International Conference on Development and Application Systems (DAS)*, 111–114.

Nhật, N. M., & Duy, N. H. K. (2024). Dự báo khả năng vỡ nợ của doanh nghiệp nhỏ và vừa tại Việt Nam: Nghiên cứu trên các mô hình học máy. *Tạp Chí Kinh Tế - Luật & Ngân Hàng*, *266*, 51–64.

Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., & Idroes, R. (2023). Credit card fraud detection for contemporary financial management using XGBoost-driven machine learning and data augmentation techniques. *Indatu Journal of Management and Accounting*, *1*(1), 29–35.

Parente, M., Rizzuti, L., & Trerotola, M. (2024). A profitable trading algorithm for cryptocurrencies using a neural network model. *Expert Systems with Applications*, *238*, 121806.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *ArXiv Preprint ArXiv:1503.06462*.

Pham, V. H. S., & Le, T. D. (2024). Research on Applying Machine Learning Models to Predict and Assess Return on Assets (Roa).

Ragb, H. (2023). Multi-layered Deep Learning Perceptron Based Model for Predicting Drug Price Changes. *Authorea Preprints*.

Rashedi, K. A., Ismail, M. T., Al Wadi, S., Serroukh, A., Alshammari, T. S., & Jaber, J. J. (2024). Multi-layer perceptron-based classification with application to outlier detection in Saudi Arabia stock returns. *Journal of Risk and Financial Management*, *17*(2), 69.

Salamzadeh, A., Ebrahimi, P., Soleimani, M., & Fekete-Farkas, M. (2022). Grocery apps and consumer purchase behavior: application of Gaussian mixture model and multi-layer perceptron algorithm. *Journal of Risk and Financial Management*, *15*(10), 424.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 553–572.

Sudhakaran, P., & Baitalik, S. (2022). XGBoost Optimized by Adaptive Tree Parzen Estimators for Credit Risk Analysis. *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 1–6.

Theil, H. (1992). Henri Theil's contributions to economics and econometrics: econometric theory and methodology. Vol. I (Vol. 1). Springer Science & Business Media.

Tsai, C.-F., & Chiou, Y.-J. (2009). Earnings management prediction: A pilot study of combining neural networks and decision trees. *Expert Systems with Applications*, *36*(3), 7183–7191.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Springer.

Tutcu, B., Kayakuş, M., Terzioğlu, M., Ünal Uyar, G. F., Talaş, H., & Yetiz, F. (2024). Predicting Financial Performance in the IT Industry with Machine Learning: ROA and ROE Analysis. *Applied Sciences*, *14*(17), 7459.

Uddin, M. S., Chi, G., Al Janabi, M. A. M., & Habib, T. (2022). Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance & Economics*, *27*(3), 3713–3729.

Wang, Y. (2022). Credit risk evaluation of asset securitization of PPP project of sports public service venues based on random forest algorithm. *Computational Intelligence and Neuroscience*, *2022*(1), 5177015.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79–82.

Wong, W. K., Cheng, Y., & Yue, M. (2024a). Could regression of stationary series be spurious? *Asia-Pacific Journal of Operational Research,* 2440017.

Wong, W. K., & Pham, M. T. (2022b). Could the test from the standard regression model could make significant regression with autoregressive noise become insignificant – a note. *The International Journal of Finance, 34*(2), 19–39.

Wong, W. K., & Pham, M. T. (2022a). Could the test from the standard regression model could make significant regression with autoregressive noise become insignificant? *The International Journal of Finance, 34*(1), 1–18.

Wong, W. K., & Pham, M. T. (2023b). Could the test from the standard regression model could make significant regression with autoregressive Yt and Xt become insignificant – a note. *The International Journal of Finance, 35*(2), 20–41.

Wong, W. K., & Pham, M. T. (2023a). Could the test from the standard regression model could make significant regression with autoregressive Yt and Xt become insignificant? *The International Journal of Finance, 35*(1), 1–19.

Wong, W. K., & Pham, M. T. (2025a). Could the correlation of a stationary series with a non-stationary series obtain meaningful outcomes? *Annals of Financial Economics*.

Wong, W. K., & Pham, M. T. (2025b). How to model a simple stationary series with a non-stationary series? *The International Journal of Finance, 37*(1), 1–19.

Wong, W. K., Pham, M. T., & Yue, M. (2024b). Could regressing a stationary series on a non-stationary series obtain meaningful outcomes – a remedy. *The International Journal of Finance, 36*(1), 1–20.

Wong, W. K., & Yue, M. (2024). Could regressing a stationary series on a non-stationary series obtain meaningful outcomes? *Annals of Financial Economics, 19*(03), 2450011.

Xia, H., An, W., & Zhang, Z. J. (2023). Credit risk models for financial fraud detection: A new outlier feature analysis method of xgboost with smote. *Journal of Database Management (JDM)*, *34*(1), 1–20.

Xie, Z., & Huang, X. (2024). A credit card fraud detection method based on mahalanobis distance hybrid sampling and random forest algorithm. *IEEE Access*.

Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications, 186*, 115716.

Zafar, M. B., & Yasin, T. (2025). Intellectual capital and financial performance of Islamic banks: a meta-analysis. *Accounting Research Journal*.

Zhao, H. (2025). Predicting Stock Prices and Optimizing Portfolios: A Random Forest and Monte Carlo-Based Approach Using NASDAQ-100. *International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, 883–892.