ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 9 ISSUE 7: 152 - 168

MACHINE LEARNING APPROACHES TO DIVIDEND PREDICTION: AN EMPIRICAL STUDY ON VIETNAM'S STOCK MARKET

Tam Phan Huy, Tuyen Vo Thi Kieu

University of Economics and Law, Ho Chi Minh City, Vietnam, and Vietnam National University, Ho Chi Minh City, Vietnam.

DOI: https://doi.org/10.71058/jodac.v9i7011

MACHINE LEARNING APPROACHES TO DIVIDEND PREDICTION: AN EMPIRICAL STUDY ON VIETNAM'S STOCK MARKET

Tam Phan Huy, Tuyen Vo Thi Kieu*

University of Economics and Law, Ho Chi Minh City, Vietnam, and Vietnam National University, Ho Chi Minh City, Vietnam.

***Corresponding author: tamph@uel.edu.vn**

Funding: The research is funded by the University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam.

Abstract: *In finance, classifying dividend payouts has garnered considerable attention due to its impact on investment decisions. While prior studies have primarily explored the legal and policy dimensions of dividend distribution, this paper centers on analyzing financial indicators to categorize dividend payouts. Most existing research has concentrated on the legal and regulatory frameworks that influence these payouts. In contrast, our study focuses exclusively on identifying the financial factors that affect dividend payout decisions. To enhance the robustness and reliability of our analysis, we implemented various data processing techniques, including managing missing values, detecting outliers, addressing data imbalances, and applying k-fold cross-validation. Our comprehensive approach involved testing 10 different models, such as KNeighbors Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, XGB Classifier, LGBM Classifier, CatBoost Classifier, ExtraTrees Classifier, AdaBoost Classifier, and Bagging Classifier. Among these, the ExtraTrees Classifier emerged as the top performer, achieving the highest accuracy at 81.61%, with a recall of 80.18%, an F1 score of 80.69%, precision of 82.52%, and Cohen's Kappa of 63.58%. The Random Forest Classifier also showed strong performance, with an accuracy of 79.33% and Cohen's Kappa of 58.58%. These findings demonstrate the efficacy of machine learning algorithms in accurately classifying dividend payouts. The resulting binary classification system is highly valuable for investors making informed decisions regarding investment opportunities, risk management, and portfolio diversification.*

Keywords: Dividend Payouts, Classification, Machine Learning, ExtraTrees Classifier, Vietnam

JEL codes: C53, E37, G17

1. Introduction

The dividend payout ratio is a critical element of corporate finance, influenced by various factors that shape corporate decision-making. This paper seeks to explore the relationship between different financial indicators and the dividend payout ratio in manufacturing companies, drawing on significant research by Affandi et al. (2019) and Baker and Wurgler (2004). Affandi et al. (2019) examined how factors such as cash ratio, debt to equity ratio (DER), receivables turnover, net profit margin (NPM), return on equity (ROE), and institutional ownership affect the dividend payout ratio of manufacturing firms on the Indonesia Stock Exchange. Their findings highlighted that receivables turnover, ROE, and institutional ownership significantly positively influenced the dividend payout ratio, while cash ratio, DER, and NPM did not. Baker and Wurgler (2004), on the other hand, introduced the catering theory of dividends, proposing that investor demand for dividend-paying stocks drives corporate dividend decisions. Their analysis showed that firms initiate or omit dividends based on prevailing investor preferences, thereby validating the catering theory as a more accurate explanation than other dividend theories.

DOI: https://doi.org/10.71058/jodac.v9i7011

Building on these foundational studies, this paper aims to deepen the understanding of the factors influencing dividend payout ratios in companies listed on the Ho Chi Minh City Stock Exchange (HOSE). By integrating the empirical insights from Affandi et al. and Baker and Wurgler, the study broadens the scope of previous research by incorporating additional variables and employing comparative analysis. Through rigorous data analysis and statistical methodologies, the paper seeks to illuminate the interaction between financial indicators and investor preferences, offering valuable insights for both practitioners and academics in corporate finance.

In addition to synthesizing existing research, this paper employs advanced data analysis techniques to identify patterns and relationships between financial indicators and dividend payout ratios within manufacturing companies. The study carefully handles data through methods such as imputing missing values, addressing outliers, and applying techniques to manage data imbalance in dividend payout classifications. K-fold cross-validation was utilized to evaluate the performance and robustness of predictive models, providing a comprehensive assessment of their generalizability. Eleven models, including logistic regression, decision trees, and XG Boost, were compared to determine which model most accurately classified dividend payouts. The study used evaluation metrics like kappa to measure the agreement between predicted and actual classifications, further enhancing the study's reliability.

Overall, this research significantly contributes to the existing literature by improving the understanding of dividend payout decisions in the manufacturing sector. The findings are instrumental for managers, investors, and policymakers, helping them make informed decisions about dividend policies and their impact on financial performance and shareholder value. By applying robust data handling techniques and rigorous modeling approaches, this study not only advances the field of dividend payout classification but also provides practical tools for navigating the complex landscape of corporate finance.

2. Literature review

2.2. Background theories

Miller and Modigliani's (1961) dividend irrelevance theory serve as the foundational concept for understanding the relationship between dividend policy, growth, and stock valuation, arguing that in perfect capital markets, dividend policy does not impact firm value, as investment policy is the primary determinant. However, when relaxing these ideal assumptions, several complementary theories have emerged to explain corporate dividend policy, including the Clientele Effect Theory, Agency Cost and Free Cash Flow Theory, and Asymmetric Information and Signaling Theory. These theories provide a comprehensive framework for analyzing factors influencing dividend payout decisions.

The Clientele Effect Theory suggests that in imperfect capital markets, investors prefer dividend policies that match their cash flow needs (Miller & Modigliani, 1961). This preference is influenced by liquidity concerns, as seen in studies by Banerjee, Gatchev, and Spindt (2007), and Brav et al. (2005), where shareholders of illiquid stocks favor dividends due to lower transaction costs. Additionally, tax preference studies (Farrar & Selwyn, 1967; Elton & Gruber, 1970) indicate that investors in higher tax brackets tend to favor lower dividend yields, further reinforcing the clientele effect. The Agency Cost and Free Cash Flow Theory, proposed by Jensen (1986), posits that dividends can mitigate agency problems by reducing excess free cash flow, which managers might otherwise invest in non-value-maximizing projects. By distributing dividends, companies align managerial actions with shareholder interests, which can lead to positive market reactions as a signal of reduced overinvestment (Jensen, 1986; Easterbrook, 1984).

DOI: https://doi.org/10.71058/jodac.v9i7011

Signaling Theory addresses the issue of information asymmetry between managers and investors, suggesting that changes in dividend policy can signal future earnings prospects (Miller & Rock, 1985). Dividend increases are generally interpreted as positive indicators of future earnings, while decreases may signal the opposite, as supported by empirical studies (Baker & Wurgler, 2004). These theories are integral to understanding the financial indicators influencing dividend payout decisions in the context of machine learning algorithms. The Clientele Effect Theory guides the analysis of investor preferences and market imperfections, Agency Cost Theory emphasizes the assessment of free cash flow and leverage to evaluate agency costs, and Signaling Theory provides a framework for interpreting the impact of dividend changes on market expectations.

2.1. Related work

The relationship between corporate governance and dividend policy has been a significant area of research, particularly in the context of Vietnamese enterprises. H. V. Nguyen et al. (2021) conducted a study examining how corporate governance impacts dividend policy in Vietnam, focusing on companies listed on the Vietnam stock exchange from 2008 to 2018, with a dataset of 2,937 observations. Utilizing GLS regression, the study found that stronger corporate governance, particularly in terms of the chairman of the board of directors (BOD) and the managing director, correlated with lower dividend payouts. Additional factors such as profitability, financial leverage, firm size, and investment opportunities were also found to influence dividend policies, highlighting the importance of both internal and external governance mechanisms in shaping income distribution strategies.

In a related study, A. H. Nguyen et al. (2021) explored the nuanced effects of dividend policies on firm performance in Vietnam. Their findings revealed that dividend payments negatively affect accounting-based performance indicators like Return on Assets (ROA) and Return on Equity (ROE), but positively influence market-based indicators such as Tobin's Q. This suggests that while higher dividends might limit reinvestment opportunities and future profitability, they enhance market perceptions of a firm's stability and reliability, highlighting the trade-offs involved in dividend policy decisions.

The literature also shows that profitability is a key determinant of dividend payments. Amidu and Abor (2006) and Fama and French (2002) suggest that firms increase dividends as profitability rises, while Jensen’s (1986) agency costs theory argues that investors demand higher dividends to prevent overinvestment when profitability is high. This positive relationship between profitability and dividend payments is supported by studies in various markets, including those by Adaoglu (2000) in Turkey and Pruitt and Gitman (1991) in the U.S., emphasizing the alignment of investment, financing, and dividend decisions with growth and investment opportunities.

Other factors influencing dividend payments include firm size, financial leverage, life cycle stage, liquidity, and risk. Larger firms with more dispersed ownership tend to pay higher dividends due to greater agency problems (Meckling & Jensen, 1976), while firms with higher leverage are likely to reduce dividends to maintain necessary cash flow for debt obligations (Crutchley & Hansen, 1989). The life-cycle theory by DeAngelo et al. (2006) also posits that firms pay lower dividends in their early stages when investment opportunities are abundant. Additionally, stock liquidity and risk play roles, with less liquid firms more likely to initiate dividends (Banerjee et al., 2007), and firms with lower risk more inclined to pay dividends as they mature (Grullon et al., 2002). Cash holdings are another factor, with studies showing a positive relationship between cash reserves and dividend payments (Li & Lie, 2006; De Cesari & Huang-Meier, 2015; Bradford et al., 2013).

DOI: https://doi.org/10.71058/jodac.v9i7011

3. Methodology

3.1. Data collection

The dataset utilized in this study was sourced from Refinitiv, comprising 5,728 records from both financial and non-financial firms listed on the Ho Chi Minh and Ha Noi stock exchanges, spanning the years 2014 to 2021. The data collection was carried out with great precision and care, ensuring the highest levels of data integrity and reliability. Strict adherence to security protocols and privacy standards was maintained throughout the process, guaranteeing that the data was collected and used in a manner that is both appropriate and responsible. The dataset includes 5,728 entries with 15 distinct variables, each representing different financial characteristics of the companies involved. Table 1 provides a detailed overview of the variables included in the dataset.

Table 1. Dataset description

Attributes	Type	Description
Symbol	Categorical	A unique series of letters representing a company on stock market
Year	Numerical	When data was recorded
Dividend Payment	Numerical	“0” means that company didn’t pay dividend, “1” means that company did pay dividend
ROA	Numerical	The ratio of a company’s profit to its total assets
ROE	Numerical	The ratio of a company’s profit to its equity capital
Free Cash Flow	Numerical	The amount of money that a company has left over after paying all costs and necessary investments to maintain existing assets
Leverage	Numerical	Leverage suggests a company’s debt level
Current ratio	Numerical	The ratio between a company’s current assets and its debt due in the near future
Company Market Capital	Numerical	The market value of a company – stock price multiplied by the total number of outstanding shares
Market to book ratio	Numerical	The ratio between a company’s market value and its book value of assets
Company age	Numerical	The number of years a company has been in operation

DOI: https://doi.org/10.71058/jodac.v9i7011

Liquidity	Numerical	Measure of a company's ability to cover short-term financial obligations
Exchange Name	Categorical	The name of the stock exchange where the company is listed
GICS Industry Name	Categorical	The name of the business industry that the company belongs to
Type	Categorical	The sector in which the business operates

Source: by author

The implementation of this study can be segmented into six distinct steps, as illustrated in Figure 1. A structured set of logical procedures will be employed to explore the hypotheses, which propose potential relationships among the variables. This research is grounded in a positivist methodology, given its primarily predictive focus. The study leverages machine learning algorithms to facilitate the attainment of its research objectives.

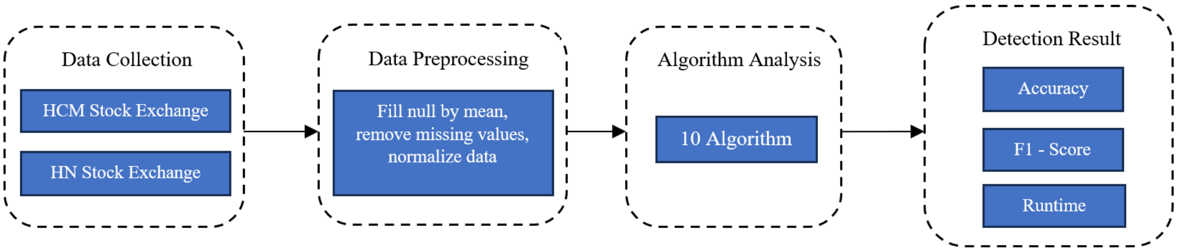

Figure 1. *The implementation process of the study*

Source: by author

3.2. Data processing procedure

3.2.1 Cleaning data

The data collected is unprocessed, often containing incomplete sections, font inconsistencies, and duplicate entries. These issues are common and expected. Our next step involves data preprocessing to refine and prepare the raw data for analysis. This process includes techniques such as data cleaning, handling missing values, removing outliers, eliminating noise, and normalizing the data. These steps are essential to enhance the accuracy, completeness, and readiness of the data for subsequent analysis.

DOI: https://doi.org/10.71058/jodac.v9i7011

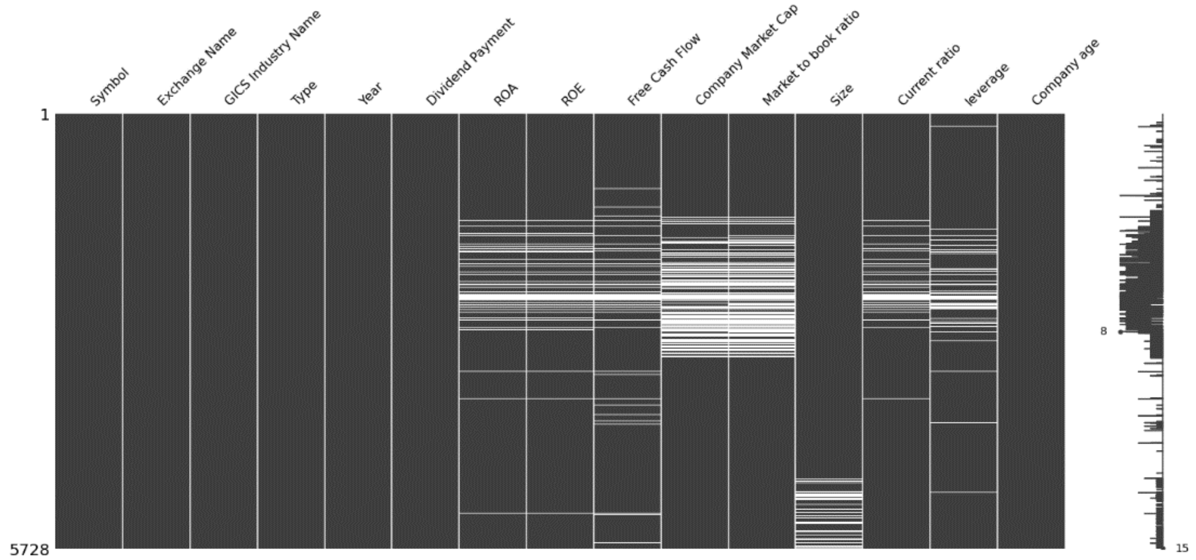

Figure 2. *The missing values of variables*

Source: by author

After excluding outliers, the dataset retains 91.3% of its original data. The correlation matrix uncovers several key relationships among the financial variables. Notably, Dividend Payment exhibits a moderate positive correlation with Company Age (0.15) and negative correlations with both Leverage (-0.26) and Size (-0.27). This indicates that older firms are more inclined to distribute dividends, whereas firms with higher leverage and larger size tend to pay less. Return on Assets (ROA) and Return on Equity (ROE) display a very strong positive correlation (0.81), suggesting that increases in one metric are closely mirrored by the other. Company Market Capitalization is strongly positively associated with Size (0.56), implying that larger firms generally possess higher market capitalizations. Leverage demonstrates a moderate negative correlation with ROA (-0.26), indicating that firms with higher asset returns typically have lower leverage. Additionally, a significant negative correlation exists between Current Ratio and Leverage (-0.26), revealing that firms with higher liquidity, as measured by the Current Ratio, tend to maintain lower leverage. Collectively, these correlations offer critical insights into the interplay of financial metrics and their impact on dividend payments, providing a foundation for further analysis and model refinement.

DOI: https://doi.org/10.71058/jodac.v9i7011

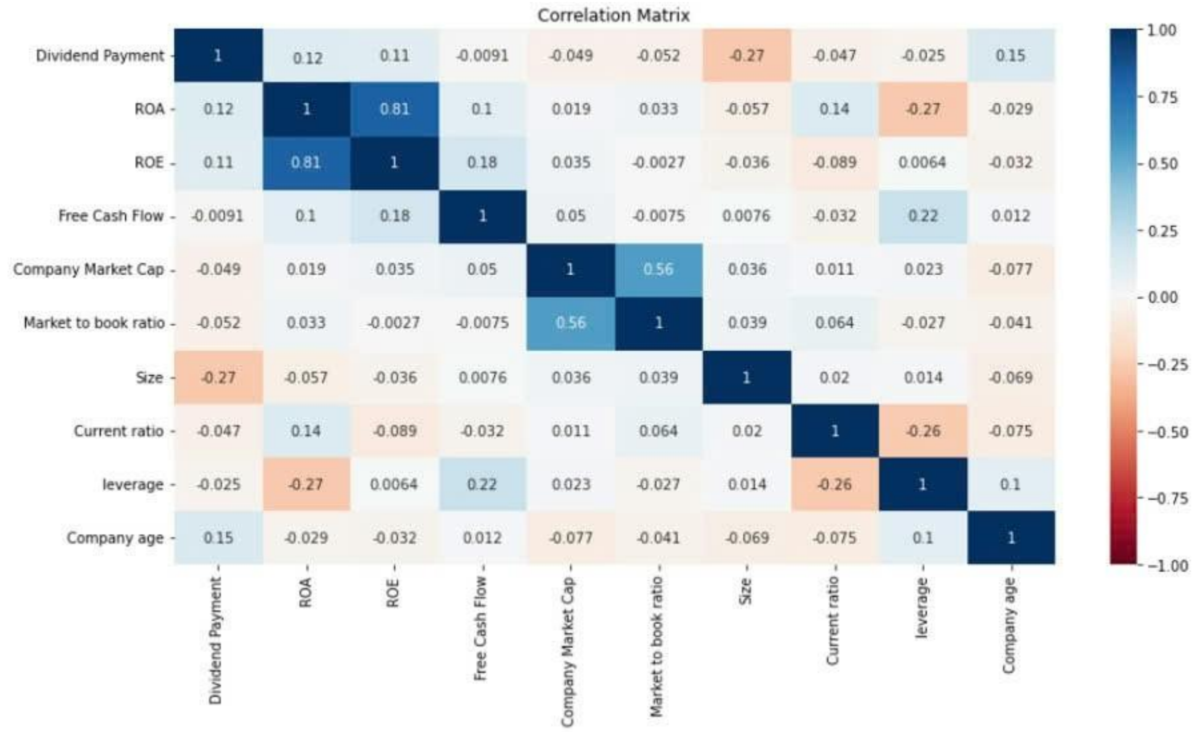

***Figure 3**. Correlation Matrix of Variables*

Source: by author

3.2.2 Data Transformation

Step 1. Transform the categorical values to numerical format

We perform numerical encoding of categorical variables because machine learning algorithms typically only work with numerical data. When we have categorical variables, they cannot be represented directly as numbers and need to be encoded into a numerical format to be used in algorithms. We use the "One-hot encoding" method to convert categorical variables into numerical format. This method converts the values of a categorical variable into dummy variables that describe the presence or absence of each value. Each dummy variable takes a value of 0 or 1, indicating the presence or absence of that value in the data.

Step 2. Normalize the data

We opted for min-max scaling as our data normalization technique, which transforms values into a specified range, typically between 0 and 1. This method is applied using the formula:

$$\text{X\_scale} = \frac{x - x\_min}{x\_\max - x\_min}$$

- x is the original value
- x_min is the minimum value in the data set
- x_max is the maximum value in the data set

This approach standardizes the units of measurement across variables, minimizing the impact of differing scales. By doing so, it facilitates fair comparisons between variables and reduces potential biases in the analysis. Additionally, min-max scaling effectively addresses outliers by compressing the range of most data points, bringing outliers closer to the upper or lower bounds of the defined range.

 DOI: https://doi.org/10.71058/jodac.v9i7011

3.3. Algorithms evaluation metric

To assess the performance and effectiveness of the machine learning models employed in this study, a comprehensive set of evaluation metrics was utilized. These metrics provide a multifaceted understanding of each model's ability to accurately classify dividend payouts, ensuring robust and reliable results. The primary evaluation metrics used include Accuracy, Precision, Recall, F1-Score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and Cohen's Kappa. Accuracy measures the proportion of correctly classified instances out of the total number of observations. While it provides a general sense of model performance, its effectiveness can be limited in cases of class imbalance.

Precision evaluates the proportion of true positive predictions relative to the total positive predictions made by the model. This metric is crucial in understanding the model's ability to minimize false positives, which is important in scenarios where the cost of incorrect positive classifications is high. Recall (or Sensitivity) assesses the proportion of actual positives that were correctly identified by the model. High recall indicates that the model effectively captures the majority of true positive cases, which is essential for ensuring that significant instances are not overlooked. F1-Score is the harmonic mean of Precision and Recall, providing a balanced measure that accounts for both false positives and false negatives. This metric is particularly useful when there is an uneven class distribution, as it offers a single performance indicator that balances the trade-offs between Precision and Recall.

4. Results and discussion

4.1. Descriptive Statistics

The study examined various financial and non-financial variables of listed firms, including ROA, ROE, Free Cash Flow, Size, Current ratio, leverage, Market to book ratio and Company Market Cap. After removing the observations with the missing value, the data was utilized, including 5728 observations with descriptive statistics as follows:

Table 3. *Descriptive Statistics*

	Count	mean	Std	min	25%	50%	75%	max
ROA	5728	6.13	7.51	-78.73	1.89	5.11	8.54	83.95
ROE	5728	11.35	14.62	-329.37	5.03	10.76	16.26	160.74
Free Cash Flow	5728	1.23e+11	1.51e+12	-4.3e+13	-2.64	1.29e+10	9.31e+10	2.72e+13
Size	5728	10.54	3.76	0.00	11.111	11.708	12.24	14.63
Current ratio	5728	4.09	12.40	0.82	1.46	1.99	3.48	374.03
leverage	5728	1.80	4.54	-135.43	0.46	1.091	1.89	176.23
Market to book ratio	5728	34.06	334.00	-24.32	0.39	1.283	20.70	15537.20
Market Cap	5728	4.69e+12	2.11e+12	3.3e+9	1.43e+11	5.36e+11	4.69e+12	3.87e+14

Source: by author

The descriptive statistics for the financial variables in the dataset reveal several important insights. With 5,728 observations per variable, the dataset is comprehensive. The mean Return on Assets (ROA) stands at 6.13%, and the mean Return on Equity (ROE) at 11.36%, indicating that, on average, the companies achieve moderate returns. However, the substantial standard

DOI: https://doi.org/10.71058/jodac.v9i7011

deviations for ROA (7.51) and ROE (14.63) suggest considerable variability across companies, with some performing exceptionally well and others poorly. Free Cash Flow exhibits an extremely high mean value of approximately 1.24e+11, but the large standard deviation (1.51e+12) and the broad range between the minimum (-4.25e+13) and maximum (2.72e+13) values highlight the presence of extreme outliers in the data, further evidenced by the negative minimum values, indicating substantial negative cash flows for some firms. The Size metric, representing the scale of companies, has a mean of 10.55 and a relatively modest standard deviation (3.77), reflecting that most companies are of similar size, though outliers exist, as indicated by the minimum value of 0 and maximum of 14.63.

The Current Ratio, with a mean of 4.10 and a significant standard deviation of 12.40, reveals wide variation in liquidity among companies, with some maintaining high liquidity and others potentially facing liquidity challenges. Leverage has a mean of 1.80 and a standard deviation of 4.55, implying that while most companies operate with moderate leverage, there are notable outliers with very high or even negative leverage, indicating financial distress or unique capital structures. The Market to Book Ratio, with a mean of 34.06 and an exceptionally large standard deviation (334.00), underscores extreme variability, suggesting that some companies are valued much higher or lower in the market relative to their book value. Finally, the Company Market Capitalization exhibits a mean of 4.69e+12 and a very large standard deviation (2.11e+13), reflecting substantial disparities in company sizes within the market, with the maximum value of 3.87e+14 indicating the presence of very large companies in the dataset.

Overall, these statistics highlight considerable variability within the dataset and the presence of outliers, especially in metrics such as Free Cash Flow and Market to Book Ratio. This variability needs to be carefully accounted for when developing and accessing predictive models.

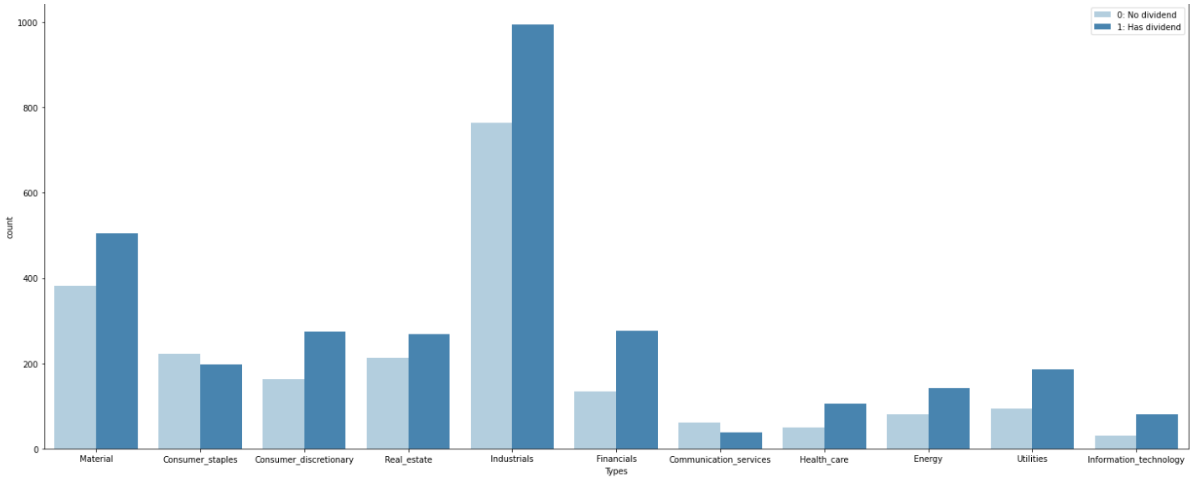

Figure 5. *Dividend Payment status in different Types*

Source: by author

The bar chart depicts the distribution of dividend payment status across various industry sectors. Sectors like Materials and Industrials demonstrate a higher propensity for companies to distribute dividends, with the Industrials sector having the largest number of dividend-paying companies. The Financials and Real Estate sectors also show a significant number of dividend-paying companies, reflecting a preference for dividend payouts. In contrast, sectors such as Consumer Staples, Consumer Discretionary, and Healthcare exhibit a more even distribution between dividend-paying and non-paying companies.

DOI: https://doi.org/10.71058/jodac.v9i7011

4.2. Results model

We'll focus on three key evaluation metrics: F1-score, Precision, and Cohen's Kappa, as these are highly regarded for assessing model performance. The F1-score offers a balanced view by considering both Precision and Recall, making it effective for evaluating a model's accuracy in classifying both positive and negative instances. Precision measures the proportion of true positive predictions among all positive predictions, crucial in contexts like medical diagnosis or fraud detection where avoiding false positives is important. Cohen's Kappa assesses the agreement between predicted and actual labels, accounting for chance agreement, and is particularly useful for imbalanced datasets, providing a more reliable measure of classification accuracy.

The 5-fold cross-validation accuracy scores reveal that the Extra Trees Classifier is the top performer, with accuracy ranging from 79.34% to 83.38% and an average of 81.60%, demonstrating its robustness and reliability in predicting both dividend payments (class 1) and non-payments (class 0). The Random Forest Classifier also performs strongly, with accuracy scores between 78.52% and 81.47% and an average of 79.33%, showing its capability to effectively manage complex data. The CatBoost Classifier follows closely with accuracy ranging from 77.16% to 80.86% and an average of 78.55%, proving its efficiency in handling categorical features and delivering consistent predictions.

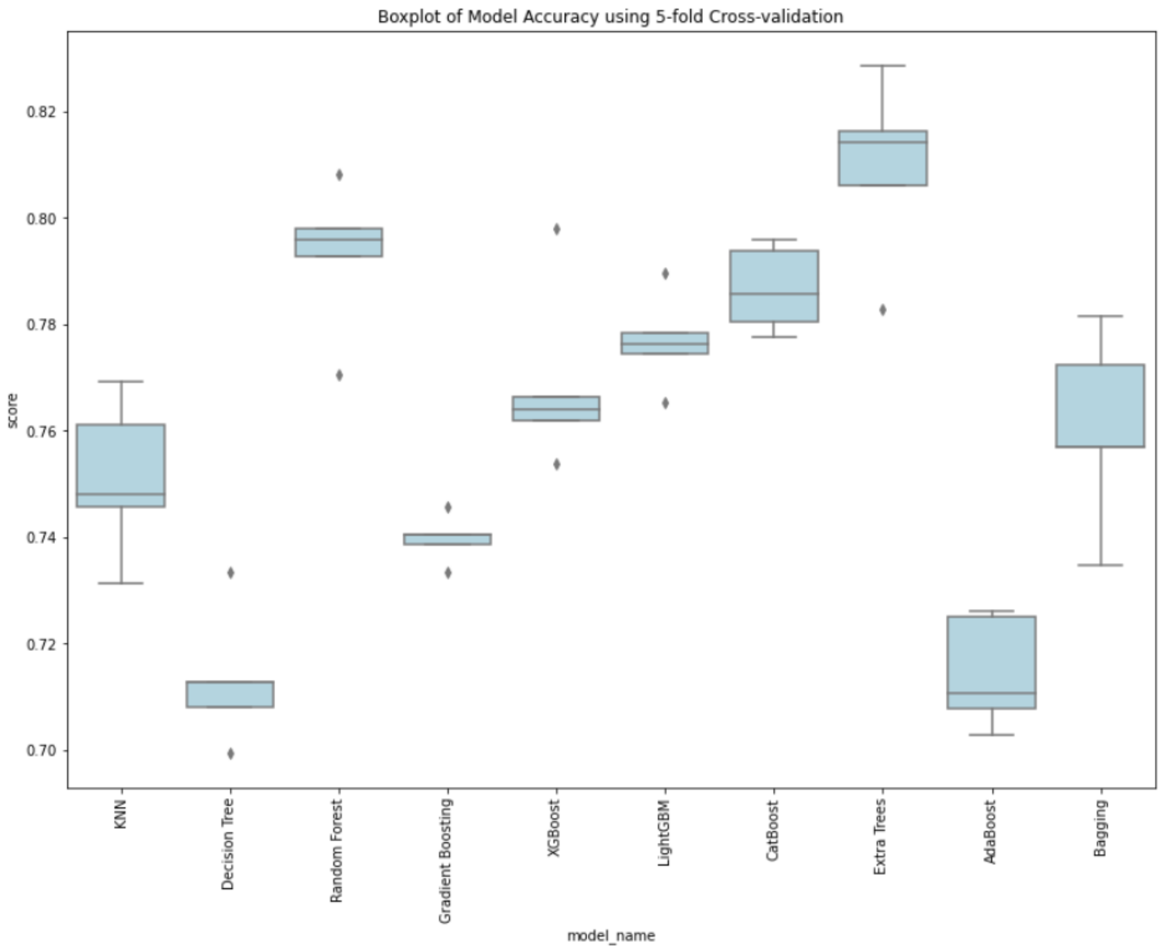

Figure 6. Boxplot of Model Accuracy using 5-fold Cross-validation

Source: by author

XGBoost also shows solid performance, with scores from 75.62% to 78.85% and an average of 77.05%, indicating balanced and effective predictions. LightGBM performs well with

DOI: https://doi.org/10.71058/jodac.v9i7011

accuracy between 76.08% and 79.75% and an average of 77.76%, highlighting its efficiency with large datasets. In contrast, KNeighborsClassifier and Bagging Classifier deliver moderate accuracy, averaging 75.44% and 76.13%, respectively. Gradient Boosting demonstrates less consistency, with an average accuracy of 73.57%. Finally, Decision Tree and AdaBoost classifiers have the lowest accuracy, averaging 71.33% and 70.93%, respectively, showing greater variability and a tendency toward overfitting. Overall, ensemble methods like Extra Trees, Random Forest, CatBoost, and XGBoost emerge as the most effective and reliable models for predicting dividend payments.

The evaluation of F1 scores using 5-fold cross-validation highlights the Extra Trees Classifier as the top performer, with consistent F1 scores between 78% and 82%, demonstrating its strong balance between precision and recall. The Random Forest and CatBoost classifiers also show robust performance, with F1 scores ranging from 76% to 79%, effectively handling complex and categorical data. XGBoost and LightGBM follow closely, delivering balanced predictions with F1 scores between 74% and 78%. Other models like KNeighborsClassifier and Bagging Classifier offer moderate performance, while Gradient Boosting, Decision Tree, and AdaBoost classifiers show lower consistency and higher variability, with F1 scores below 76%.

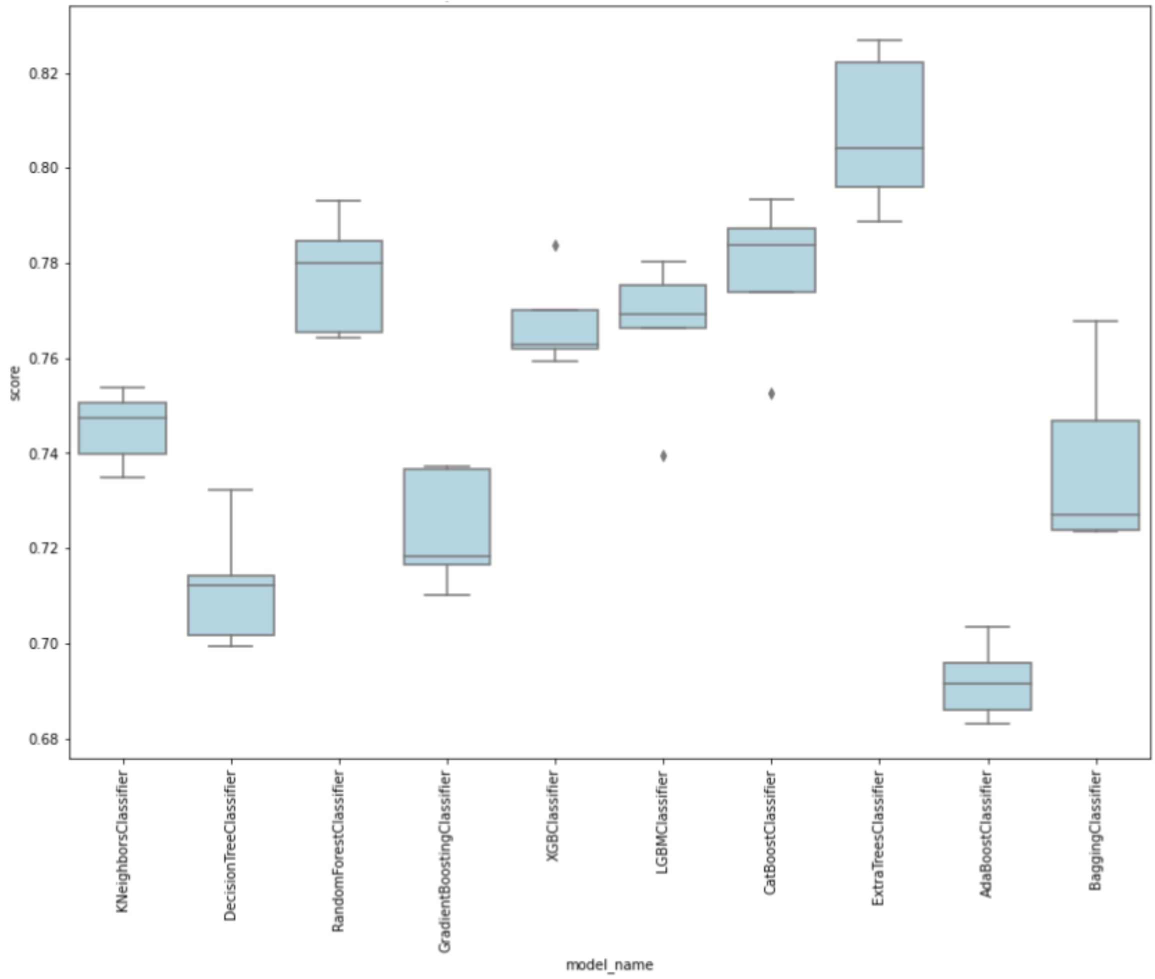

Figure 7. Boxplot of Model F1-Score using 5-fold Cross-validation

Source: by author

Detailed performance metrics further confirm the superiority of the Extra Trees Classifier, which leads in accuracy (81.61%), recall (80.18%), F1 score (80.69%), precision (82.52%), and Cohen's Kappa (63.58%). Random Forest, CatBoost, XGBoost, and LightGBM also

DOI: https://doi.org/10.71058/jodac.v9i7011

perform strongly, with high accuracy and reliable predictions, making them the most effective models for predicting dividend payments. In contrast, models like Decision Tree and AdaBoost exhibit lower performance, with a tendency to overfit and less reliable results. Overall, the Extra Trees, Random Forest, CatBoost, XGBoost, and LightGBM classifiers stand out as the best options for this task.

Table 4. *Mean value metrics of each model*

Model name	Accuracy	Recall	F1	Precision	Cohen's Kappa
ExtraTreesClassifier	0.754430	0.734486	0.749421	0.765515	0.508866
DecisionTreeClassifier	0.713334	0.696201	0.710607	0.719185	0.435926
RandomForestClassifier	0.793307	0.785229	0.792299	0.807428	0.585812
GradientBoostingClassifier	0.735697	0.749386	0.738196	0.728960	0.470187
XGBClassifier	0.770545	0.765496	0.769230	0.773416	0.541086
LGBMClassifier	0.777595	0.778792	0.777723	0.777013	0.555192
CatBoostClassifier	0.785452	0.783627	0.784916	0.786489	0.570904
ExtraTreesClassifier	0.816071	0.801758	0.806932	0.825221	0.635766
AdaBoostClassifier	0.709305	0.723194	0.713145	0.703557	0.418605
BaggingClassifier	0.761275	0.704251	0.751910	0.783128	0.523356

Source: by author

Besides considering those metrics, there is a need to consider the running time of the models to choose which model is best to predict the ability to pay dividends.

DOI: https://doi.org/10.71058/jodac.v9i7011

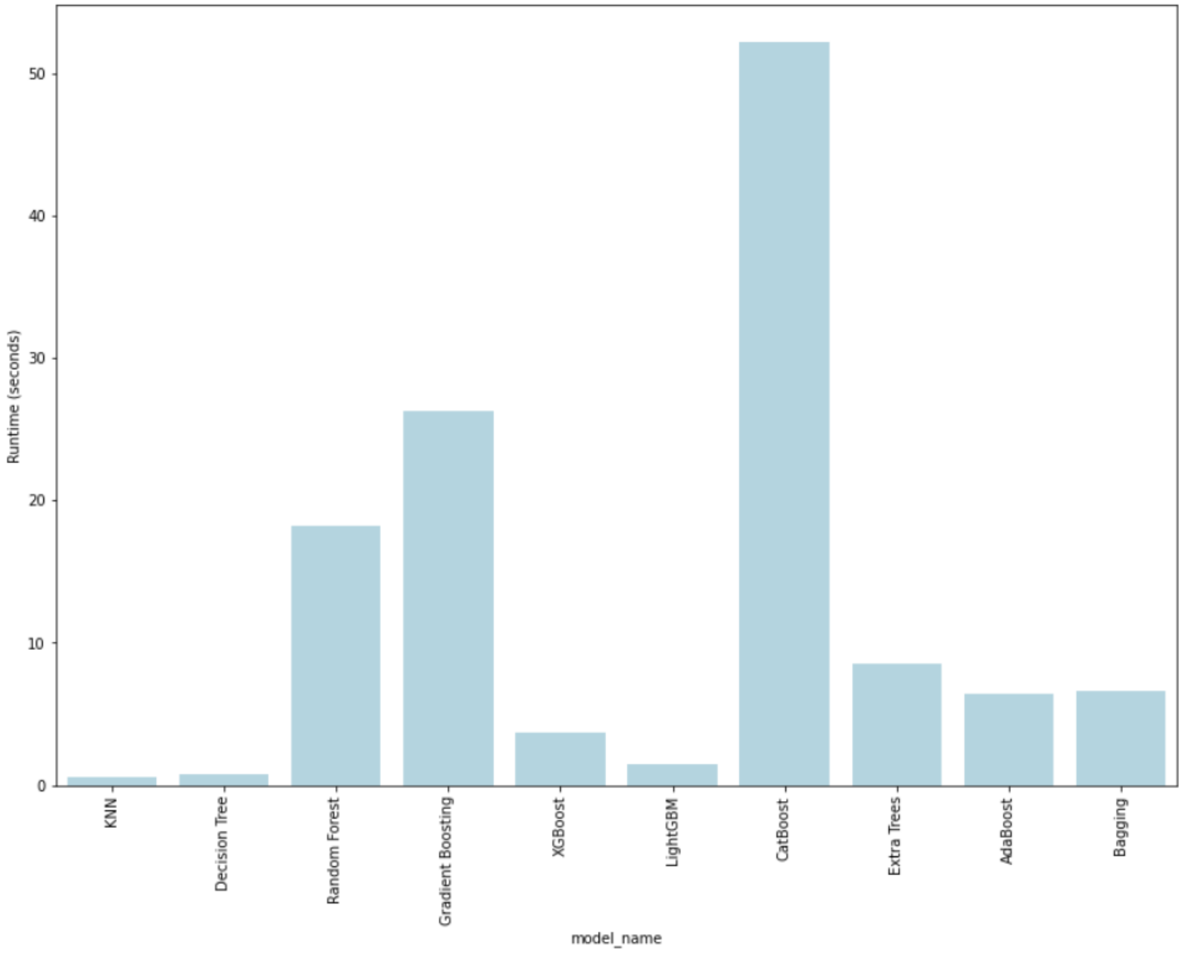

Fig 8. The Runtime of each model

Source: by author

The runtime comparison of the models, as depicted in the bar chart, offers key insights into their computational efficiency for predicting dividend payments. The CatBoost Classifier has the longest runtime, exceeding 50 seconds, indicating its high computational demand. Gradient Boosting follows with a runtime of around 30 seconds. Models like Random Forest, Extra Trees, AdaBoost, and Bagging Classifiers have moderate runtimes, ranging from approximately 15 to 25 seconds. On the other hand, KNeighborsClassifier, Decision Tree, XGBoost, and LightGBM have the shortest runtimes, all under 10 seconds, with KNeighborsClassifier and Decision Tree being the fastest.

When considering both performance metrics and runtime efficiency, the Extra Trees Classifier stands out as the most suitable model for predicting dividend payments. It achieved the highest accuracy (81.61%), recall (80.18%), F1 score (80.69%), precision (82.52%), and Cohen's Kappa (63.58%), showcasing its superior predictive power and generalizability. Its consistent performance across multiple metrics underscores its reliability in accurately classifying companies that will pay dividends (class 1) and those that will not.

The Random Forest Classifier also demonstrated strong performance, with an accuracy of 79.33%, recall of 78.52%, F1 score of 79.23%, precision of 80.74%, and Cohen's Kappa of 58.58%. This highlights its effectiveness in managing complex datasets and delivering accurate predictions, making it a solid choice for practical applications. Both CatBoost and LightGBM also performed well, with CatBoost achieving an accuracy of 78.55%, recall of 78.36%, F1 score of 78.49%, precision of 78.65%, and Cohen's Kappa of 57.09%. LightGBM recorded an

DOI: https://doi.org/10.71058/jodac.v9i7011

accuracy of 77.76%, recall of 77.88%, F1 score of 77.77%, precision of 77.70%, and Cohen's Kappa of 55.52%. These models strike a good balance between predictive accuracy and computational efficiency, making them valuable tools for dividend prediction.

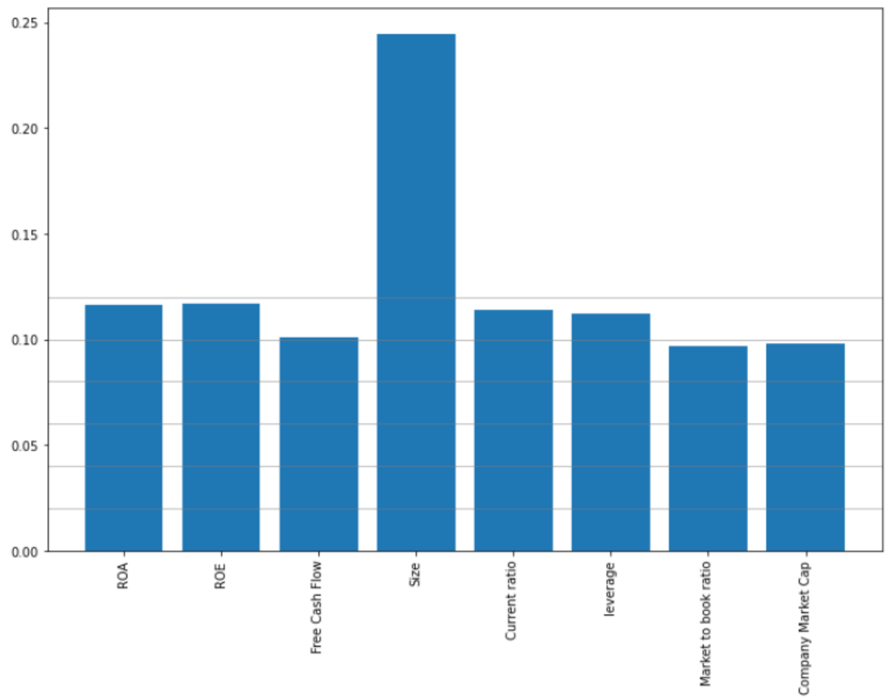

Figure 9. *Feature Importances – Extra Trees Model*

Source: by author

XGBoost demonstrated solid performance with an accuracy of 77.05%, recall of 76.55%, F1 score of 76.92%, precision of 77.34%, and Cohen's Kappa of 54.11%. Although it didn't surpass the leading models, its balanced predictions make it a dependable option for diverse datasets. KNeighbors Classifier and Bagging Classifier delivered moderate performance, with average accuracies of 75.44% and 76.13%, respectively. Gradient Boosting showed less consistency, with an average accuracy of 73.57%, making these models less optimal than the top performers but still useful in certain scenarios. Decision Tree and AdaBoost classifiers had the lowest performance, with accuracies of 71.33% and 70.93%, respectively, indicating high variability and a tendency to overfit, which limits their effectiveness for accurate dividend predictions.

Regarding runtime, the CatBoost Classifier was the most time-consuming, taking over 50 seconds, reflecting significant computational demands. Gradient Boosting followed at around 30 seconds. Random Forest, Extra Trees, AdaBoost, and Bagging Classifiers had moderate runtimes between 15 and 25 seconds. KNeighborsClassifier, Decision Tree, XGBoost, and LightGBM had the shortest runtimes, all under 10 seconds, with KNeighborsClassifier and Decision Tree being the quickest.

In summary, the Extra Trees Classifier emerges as the top model, offering high accuracy, consistent F1 scores, and reasonable computational efficiency. The Random Forest Classifier is a strong second, providing robust performance and good generalization. While CatBoost and Gradient Boosting deliver strong results, their longer runtimes may be a disadvantage in time-sensitive contexts. XGBoost and LightGBM offer a good balance between performance and computational efficiency but are slightly less effective than the top performers.

DOI: https://doi.org/10.71058/jodac.v9i7011

5. Conclusion & recommendations

This paper investigates the relationship between various financial indicators and the dividend payout ratio within manufacturing companies, drawing on two key research studies: Affandi et al. (2019) on the impact of financial ratios and institutional ownership on dividend payout, and Baker and Wurgler's (2004) "A Catering Theory of Dividends." To enhance the precision and reliability of the findings, advanced data analysis techniques were incorporated throughout the research process.

This study contributes to finance literature by evaluating the application of machine learning algorithms in predicting dividend payouts for firms listed on the Vietnam stock exchange. It advances existing knowledge on dividend policies and offers a practical tool for companies to forecast dividends and devise effective management strategies. The findings underscore the significance of machine learning in predicting dividend payments, a complex financial decision, and provide essential insights for financial decision-making. The study aimed to assess the performance of various classification algorithms in forecasting dividend payments, offering comprehensive insights for financial decision-making. Results indicate that while models like Random Forest and CatBoost perform well in terms of accuracy and F1 scores, the Extra Trees Classifier emerges as the most effective algorithm, demonstrating the highest accuracy and consistent performance, particularly in identifying dividend-paying firms (class 1). The alignment of these results with prior research strengthens the robustness and validity of the study.

The study has significant implications for researchers and practitioners interested in understanding the factors influencing dividend payments. The application of machine learning models, particularly the Extra Trees Classifier, provides valuable insights into the financial decision-making processes of firms. Companies are advised to regularly monitor financial metrics such as company size, ROA, and ROE to ensure effective dividend policies and optimal resource allocation. Investors and analysts can leverage the classification results from this study to identify firms likely to pay dividends, thereby making informed investment decisions. Companies identified as potential dividend payers can conduct thorough internal reviews and refine their financial strategies to sustain or enhance dividend distributions.

However, the study has limitations, primarily focusing on firms listed on the Vietnam stock exchange and their attributes from 2022 to 2024. Future research could extend the timeframe and incorporate non-financial variables such as organizational culture, innovation, and technology. The study's focus on quantitative factors excludes qualitative aspects like company reputation, brand image, and management quality, which may also influence dividend payout ratios and should be considered in future research. Additionally, external factors like industry competition, tax policies, and other socio-economic influences that might affect a company's dividend decisions were not accounted for. The research is limited to examining the relationship between financial indicators and dividend payout ratios within a specific period, without considering past or future changes in dividend policies, which may affect the practical applicability of the study's findings. Expanding the scope to include manufacturing companies from various cities and industries could test the generalizability of the results. Moreover, applying different analytical methods, such as nonlinear regression analysis and time series analysis, could yield alternative insights. Future studies could also consider additional factors such as industry financial conditions, broader economic conditions, and explore the relationships between dividend decisions and other financial indicators like profit, assets, and liabilities. Furthermore, examining the impact of fiscal and financial policies on company dividend decisions would provide a more comprehensive understanding.

Funding

DOI: https://doi.org/10.71058/jodac.v9i7011

The research is funded by the University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam.

References

Adaoglu, C. (2000). Instability in the dividend policy of the Istanbul Stock Exchange (ISE) corporations: evidence from an emerging market. Emerging Markets Review, 1(3), 252-270.

Affandi, F., Sunarko, B., & Yunanto, A. (2019). The impact of cash ratio, debt to equity ratio, receivables turnover, net profit margin, return on equity, and institutional ownership to dividend payout ratio. Journal of Research in Management, 1(4), 1-11.

Agrawal, A., & Jayaraman, N. (1994). The dividend policies of all-equity firms: A direct test of the free cash flow theory. Managerial and decision economics, 15(2), 139-148.

Amidu, M., & Abor, J. (2006). Determinants of dividend payout ratios in Ghana. The journal of risk finance, 7(2), 136-145.

Baker, M., & Wurgler, J. (2004). A catering theory of dividends. The Journal of finance, 59(3), 1125-1165.

Banerjee, S., Gatchev, V. A., & Spindt, P. A. (2007). Stock market liquidity and firm dividend policy. Journal of Financial and Quantitative Analysis, 42(2), 369-397.

Banyi, M. L., & Kahle, K. M. (2014). Declining propensity to pay? A re-examination of the lifecycle theory. Journal of Corporate Finance, 27, 345-366.

Bradford, W., Chen, C., & Zhu, S. (2013). Cash dividend policy, corporate pyramids, and ownership structure: Evidence from China. International Review of Economics & Finance, 27, 445-464.

Bulan, L., Subramanian, N., & Tanlu, L. (2007). On the timing of dividend initiations. Financial management, 31-65.

Crutchley, C. E., & Hansen, R. S. (1989). A test of the agency theory of managerial ownership, corporate leverage, and corporate dividends. Financial management, 36-46.

De Cesari, A., & Huang-Meier, W. (2015). Dividend changes and stock price informativeness. Journal of Corporate Finance, 35, 1-17.

DeAngelo, H., DeAngelo, L., & Stulz, R. M. (2006). Dividend policy and the earned/contributed capital mix: a test of the life-cycle theory. Journal of financial economics, 81(2), 227-254.

Faccio, M., Lang, L. H. P., & Young, L. (2001). Dividends and expropriation. American economic review, 91(1), 54-78.

Fairchild, R., Guney, Y., & Thanatawee, Y. (2014). Corporate dividend policy in Thailand: Theory and evidence. International Review of Financial Analysis, 31, 129-151.

Fama, E. F., & French, K. R. (2002). Testing trade-off and pecking order predictions about dividends and debt. Review of financial studies, 1-33.

Gaver, J. J., & Gaver, K. M. (1993). Additional evidence on the association between the investment opportunity set and corporate financing, dividend, and compensation policies. Journal of Accounting and economics, 16(1-3), 125-160.

Grullon, G., Michaely, R., & Swaminathan, B. (2002). Are dividend changes a sign of firm maturity? The journal of Business, 75(3), 387-424.

DOI: https://doi.org/10.71058/jodac.v9i7011

Gugler, K., & Yurtoglu, B. B. (2003). Corporate governance and dividend pay-out policy in Germany. European economic review, 47(4), 731-758.

Li, W., & Lie, E. (2006). Dividend changes and catering incentives. Journal of financial economics, 80(2), 293-308.

Manos, R., Murinde, V., & Green, C. J. (2012). Dividend policy and business groups: Evidence from Indian firms. International Review of Economics & Finance, 21(1), 42-56.

Meckling, W. H., & Jensen, M. C. (1976). Theory of the Firm. Managerial Behavior, Agency Costs and Ownership Structure.

Nguyen, A. H., Pham, C. D., Doan, N. T., Ta, T. T., Nguyen, H. T., & Truong, T. V. (2021). The effect of dividend payment on firm's financial performance: An empirical study of Vietnam. Journal of Risk and Financial Management, 14(8), 353.

Nguyen, H. V., Dang, H. N., & Dau, H. H. (2021). Influence of corporate governance on dividend policy in Vietnam. The Journal of Asian Finance, Economics and Business, 8(2), 893-902.

Pruitt, S. W., & Gitman, L. J. (1991). The interactions between the investment, financing, and dividend decisions of major US firms. Financial review, 26(3), 409-430.

Smith Jr, C. W., & Watts, R. L. (1992). The investment opportunity set and corporate financing, dividend, and compensation policies. Journal of financial economics, 32(3), 263-292.