Science & Technology Development Journal – Economics - Law and Management 2024, ():1-12

Open Access Full Text Article

Research Article

Predicting stock price trends by machine learning of listed companies on the Ho Chi Minh City Stock Exchange

Tam Phan Huy[1,2,*], Dieu Doan Thi Ngoc[1,2]

ABSTRACT

This research explores the potential of machine learning techniques to forecast stock price trends of entities on the Ho Chi Minh City Stock Exchange, focusing on non-banking, insurance, and securities sectors. The study spans seven years, from 2015 to 2022, scrutinizing historical stock data. By implementing advanced machine learning algorithms like Support Vector Classification, Logistic Regression, and Random Forest, the research aims to determine the most effective method for accurate trend prediction. The findings are significant, revealing that the Random Forest algorithm outperforms others, offering a balanced approach in precision and recall rates. This insight is crucial for investors and financial analysts in making informed decisions, especially in the context of a developing and dynamic market like Vietnam. The research underscores the power of machine learning in financial forecasting, highlighting its potential to revolutionize investment strategies. The study's conclusion emphasizes the importance of integrating machine learning tools, particularly Random Forest, in financial analysis and decision-making processes. This research not only offers a practical tool for investors but also contributes significantly to the academic literature on financial market predictions using machine learning methodologies.

Key words: stock trend prediction, machine learning, financial ratios

[1]*University of Economics and Law, Ho Chi Minh City, Vietnam.*

[2]*Vietnam National University, Ho Chi Minh City, Vietnam.*

Correspondence

Tam Phan Huy, University of Economics and Law, Ho Chi Minh City, Vietnam.

Vietnam National University, Ho Chi Minh City, Vietnam.

Email: tamph@uel.edu.vn

History

- Received: 17-1-2024
- Revised: 30-5-2024
- Accepted: 15-8-2024
- Published Online:

DOI :

Check for updates

INTRODUCTION

In the era of rising digital technology and the emergence of artificial intelligence with connectivity, advanced analytical techniques, and automation, humans have endeavored to apply these technological achievements to the fields of economics, finance, and life in modern society. Among these, the most notable is the practical application of artificial intelligence technology in general and the machine learning branch in particular in the field of financial investment. Specifically, learning and applying machine learning in artificial intelligence has become one of the foundations for predictive tools and investment decision recommendation systems. However, because it is a potential market with thousands of new investors wishing to participate, the fact that these investors are inexperienced and may act on emotion or follow crowd psychology can lead them to make wrong decisions or lose their inherent trust in a promising market. Accurate decisions are largely based on fundamental and technical analysis skills, providing useful information and logical, reliable, and directed choices.

Researchers have highlighted the feasibility and effectiveness of using a branch of artificial intelligence, in this case, machine learning, combined with fundamental analysis of company indices on the stock exchange to analyze and assess potential price trends, and to provide a basis for investors to make logical, observant, and risk-limited decisions [1–5]. This involves using various machine learning algorithms to study and process historical data from many stock exchanges, showing great potential in making accurate predictions or assisting in the analysis of fundamental indices, making reinforced decisions based on these factors. From both practical and theoretical perspectives, the application of Machine Learning can replace human factors in automating the "learning" process and analyzing vast amounts of data with almost absolute accuracy, while also minimizing mistakes that can be made by humans. Unlike humans, machine learning can thoroughly process information and data regardless of size or scrutinize the smallest fluctuations, factors that humans might inadvertently overlook, to produce results that most accurately reflect the intrinsic values of a company when combined with fundamental analysis [6].

The primary objective of this research is to explore and validate the effectiveness of machine learning techniques in predicting the price trends of listed companies on the Ho Chi Minh City Stock Exchange. This study aims to bridge the gap between advanced

Cite this article : Huy T P, Ngoc D D T. **Predicting stock price trends by machine learning of listed companies on the Ho Chi Minh City Stock Exchange**. *Sci. Tech. Dev. J. - Eco. Law Manag.* 2024; ():1-12.

computational methods and practical financial investment strategies by leveraging the predictive power of machine learning algorithms. The significance of this research lies in its potential to enhance the decision-making process for investors, particularly those new to the market, by providing more accurate, data-driven insights. By integrating machine learning with fundamental analysis of financial data, the study seeks to offer a robust tool that can help in mitigating investment risks and maximizing returns. This approach is especially crucial in the context of a rapidly evolving and increasingly complex financial market, where traditional methods of analysis may fall short. The research is set to contribute significantly to the field of financial technology, offering a novel perspective on how artificial intelligence can revolutionize investment strategies and market analysis, ultimately democratizing access to sophisticated investment tools for a broader range of investors.

The research is organized into five chapters, beginning with an "Introduction" that provides an overview of Vietnam's economic background and stock market, along with the study's objectives, scope, and methods. The second chapter delves into the "Literature Review" discussing machine learning and algorithms like Random Forest, SVC, and Logistic Regression, and reviews existing literature. "Methodology," the third chapter, describes the research process, data collection, and variables used. The fourth chapter presents the "Results & Discussion," analyzing the predictive model's accuracy and precision. Finally, the "Conclusion and Recommendations" chapter evaluates the model's stability and suggests future research directions and data enhancements. This structure aims to comprehensively explore and validate the application of machine learning in stock market prediction.

LITERATURE REVIEW

Background theories

In this research, a comprehensive analysis of the stock market dynamics and investor behavior, particularly within the Vietnamese context, necessitates an integrated approach to financial theories. Behavioral Finance Theory, the Efficient Market Hypothesis (EMH), and Prospect Theory collectively offer a multifaceted view of market behavior, each providing unique insights into investor decision-making and market efficiency. Behavioral Finance delves into the psychological aspects of financial decision-making, highlighting how cognitive biases and emotional reactions often drive investor behavior, leading to potential market inefficiencies [7]. This stands in contrast to the principles of EMH, which assert that stock prices efficiently reflect all available information, suggesting that achieving returns above average market performance is unlikely due to the market's rational response to information [8].

The juxtaposition of Behavioral Finance and EMH presents a fundamental debate in financial theory: do markets efficiently reflect rational valuation of information, as EMH suggests, or are they frequently the result of irrational investor behaviors influenced by psychological factors? This debate is further enriched by the incorporation of Prospect Theory [9]. Prospect Theory posits that investors exhibit loss aversion, where their responses to losses are more intense than to equivalent gains. This concept provides a psychological foundation for the deviations from rational decision-making observed in Behavioral Finance and challenges the EMH assumption of rational actors in the market. The theory is particularly relevant in explaining market phenomena such as overreactions or underreactions to new information, which lead to price movements deviating from fundamental values. The application of these theories in empirical research has often focused on market anomalies that traditional financial models struggle to explain. For instance, studies like those by Baker and Wurgler (2007) [10] have utilized Behavioral Finance and Prospect Theory to elucidate the impact of investor sentiment on stock returns, offering explanations for why stocks might be overvalued or undervalued in specific scenarios. Similarly, EMH has been the subject of numerous studies testing its validity, especially in emerging markets like Vietnam, where market efficiency might differ from that in developed markets [11]. These studies underscore the complexity of market dynamics and the multifaceted nature of investor behavior, which your research aims to unpack further.

In the Vietnamese stock market, these theories can be instrumental in understanding its unique characteristics and behaviors. Behavioral Finance can shed light on how local investor biases and cultural factors influence market trends, while EMH offers a counterpoint by suggesting the influence of global market information and rational analysis on stock prices. Prospect Theory adds a layer of understanding by examining how Vietnamese investors might react differently to gains and losses, potentially driving market volatility. Through this lens, your research could explore specific phenomena such as the prevalence of herding behavior, the impact of news on stock prices, and discrepancies between stock prices and underlying fundamental values. This integrative approach not only contributes to academic discourse but also provides practical insights for investors and policymakers in navigating the complexities of the Vietnamese stock market.

Empirical research

The integration of machine learning in stock market prediction represents a significant shift in financial analysis. Machine learning techniques vary widely, from traditional algorithms to advanced deep learning models. Patel et al. (2015)[12] and Nabipour, Nayyeri, Jabani, Mosavi and Salwana (2020)[13] compared machine learning models like Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, and Naive Bayes, using technical indicators as model inputs. Their studies, focusing on accuracy metrics like RMSE and MAPE, suggest a nuanced effectiveness of these models based on dataset characteristics. Similarly, Vijh et al. (2020)[5] and Shen and Shafiq (2020)[14], explored the effectiveness of SVM, Random Forest, KNN, and Naive Bayes. Vijh et al. (2020)[5] findings suggested Random Forest's superior performance for larger datasets, a finding echoed by Shen and Shafiq (2020)[14], who noted a general preference for Random Forest in complex data scenarios. On the other hand, Khoa and Huynh (2022) [17] highlighted the exceptional accuracy (92.48%) of SVM in predicting the VN30 index, underscoring SVM's robustness in certain market conditions.

In contrast, deep learning approaches, as explored by Shen and Shafiq (2020)[14] and Wang, Fan and Wang (2021)[15], demonstrate the evolving landscape of machine learning in stock prediction. Shen and Shafiq (2020)[14] demonstrate comprehensive deep learning system, which included extensive feature engineering, outperformed traditional machine learning models, indicating the potential of deep learning in handling the complexities of stock market data. Wang, Fan and Wang (2021)[15] also observed the superior performance of deep learning methods over traditional machine learning techniques like SVM and Random Forest. Additionally, Ngoc Hai et al. (2020)[16] examine different LSTM architectures for the Vietnamese stock market revealed Bidirectional LSTM's accuracy, showcasing the effectiveness of specific deep learning architectures in certain market contexts.

The previous studies collectively indicate a growing preference for deep learning models, especially in markets with large and complex datasets. However, traditional machine learning models like Random Forest and SVM continue to hold significance, particularly in specific market conditions or when analyzing certain financial indices. The regional focus of these studies, especially on emerging markets like Vietnam and India, provides crucial insights. The studies by Patel et al. (2015)[12], Vijh et al. (2020)[5], Ngoc Hai et al. (2020)[16] and Khoa & Huynh, (2022)[17] illustrate the varied effectiveness of machine learning models in these markets, reflecting regional market idiosyncrasies. This regional specificity is crucial in understanding the global applicability of machine learning in stock prediction.

Moreover, the application of machine learning in predicting specific sectors or indices, as seen in studies by Nabipour, Nayyeri, Jabani, Shahab and Mosavi (2020)[4] and Khiem et al. (2021)[18], demonstrates the versatility of machine learning models. These studies indicate the potential of machine learning in diverse market segments, from petroleum and metals to Vietnamese shrimp prices. The integration of financial news in models by Huynh, Dang and Duong (2017)[19] and Le Hong et al. (2022)[20] for predicting the VN30 Index, further illustrates the innovative use of non-traditional data in enhancing model accuracy. These studies highlight the dynamic evolution of machine learning in stock market prediction, with a gradual shift from traditional algorithms to more complex deep learning models. The effectiveness of these models varies based on dataset characteristics, market conditions, and regional specificities. The integration of diverse data types, including financial news, indicates a broader trend towards comprehensive, multi-faceted predictive models. While deep learning shows promising results, traditional machine learning models remain relevant in certain contexts. Future research could benefit from exploring the integration of macroeconomic factors and global market trends, potentially enhancing the predictive accuracy and robustness of these models in the volatile domain of stock market prediction.

METHODOLOGY

Data

The research employed a comprehensive dataset encompassing the period from 2015 to 2022, focusing on the historical closing price data of 364 companies. These companies, excluding those from the Banking, Insurance, and Securities sectors, are listed on the Ho Chi Minh City Stock Exchange (HOSE). The dataset, which comprises 2799 observations, was derived from Refinitiv, a provider of secondary data sources. It includes 19 variables categorized into five groups based on financial indices extracted from quarterly financial reports over several years.

The study's target variable is based on the closing stock prices, defined by two conditions: a stock is labeled as 1 (indicating an upward trend) if the closing price at time t+1 is greater than the closing price at time t. Conversely, a stock is labeled as 0 (indicating a stagnant or downward trend) if the closing price at time t+1 is less than or equal to the closing price at time t. This approach to labeling provides a clear framework for the study's predictive modeling, aiming to forecast stock price movements, a critical factor for investors' decision-making processes.

Input Variables for Machine learning Algorithms

This research draws on the foundational work of three key studies in the field. The first, conducted by Christian S. in 2015, explored the impact of valuation indices on stock price changes for retail companies on the Indonesian stock exchange. The study introduced and employed variables such as the Price-to-Earnings (P/E) ratio, Dividend Yield (D/Y), Earnings Per Share (EPS), and Book to Market ratio. These variables were posited to have a relationship with stock returns. Employing various statistical tests like the Park Test, Run Test, Multicollinearity Test, and Kolmogorov-Smirnov (K-S) Test, and incorporating data transformations, Christian S. suggested that other variables, including firm size and cash flow, might also influence stock outcomes. The research highlighted the significant roles of D/Y, P/E, and the Book to Market ratio in predicting stock consequences. However, it also acknowledged limitations due to the constrained number of input indices used. This limitation contributed to the less effective and less reliable outcomes of the R square value. The study's scope was further limited by its sample size and the fact that it relied on data from 2011 to 2013. Consequently, it proposed incorporating additional potential variables such as net profit margin, return on equity, total assets turnover (TATO), and market to book ratio, as well as firm size and cash flow, in future research based on more recent data.

The second study, conducted by Naknok (2022)[21], investigated the operational efficiency of 100 Thai companies during the 2016 - 2020 period, which included the onset of the COVID-19 pandemic. This research utilized key indicators like total assets turnover (TATO), P/E ratio, Book to Market (B/M) ratio, interest coverage ratio (INTE ratio), and firm size (SIZE) to calculate corporate efficiency. The data reflected the quality of the businesses, with dependent variables including EPS and Return on Equity (ROE). Both studies collectively contribute to a comprehensive understanding of the factors influencing stock performance and corporate efficiency, providing a robust foundation for further empirical analysis in this field.

Finally, the research conducted by Dimitrantzou, Psomas and Vouzas (2023)[22] on fundamental analysis techniques in the Food and Beverage (F&B) sector significantly contributes to the existing body of work, particularly in the use of financial indices to understand their impact on stock prices. The authors employed and validated several key financial metrics including Debt to Equity ratio (D/E), Return on Asset (ROA), Current Ratio (CR), Price to Earnings (P/E) ratio, and Total Asset Turnover (TATO), to establish their relationship with stock price fluctuations.

Building upon the insights and methodologies of previous studies, the authors have developed a comprehensive table of financial indices categorized into groups of financial ratios, serving as key input coefficients. This innovative approach involved expanding the scope of their research to include a total of 19 financial indices including: (i) Cash, (ii) Cash from operating activity, (iii) Cash from Investing Activities, (iv) Cash from Financing Activities (v) Asset Turnover, (vi) Current Liabilities, (vii) Current Asset, (viii) Total Asset, (ix) Total Liabilities, (x) EPS, (xi) D/E, (xii) ROA, (xiii) Net Margin, (xiv) Revenue, (xv) Net Profit, (xvi) Current Ratio, (xvii) P/E, (xviii) P/B, (xix) Book Value Per Share, align with the research of as Table 1.

This expansion not only provides a broader data set but also equips machine learning algorithms with the necessary information to operate with higher efficiency and acceptable accuracy levels. The incorporation of these indices is a testament to the evolving nature of financial research and highlights the importance of a detailed and multi-faceted approach in understanding stock market dynamics.

Algorithm specifications

The selection of machine learning algorithms, namely Support Vector Classification (SVC), Logistic Regression, and Random Forest, in this research is driven by their specific strengths and applicability in the context of the research topic. Each of these algorithms offers distinct advantages that align with the research objectives.

Support Vector Classification (SVC) is a powerful machine learning algorithm used for binary classification tasks[34]. It works by finding the optimal hyperplane that best separates data points into distinct classes. In the context of our research on stock market trend prediction, SVC is particularly suitable due to its capacity to discern intricate boundaries between upward

Table 1: Input Variables for Machine Leaning Algorithms

No.	Variables	References
1	Cash	Milosevic (2016)[23], Amel-Zadeh et al. (2020)[24], Rouf et al. (2021)[25], Hu, Zhao and Khushi (2021)[26], Kotios et al. (2022)[27]
2	Cash from operating activity	
3	Cash from Investing Activities	
4	Cash from Financing Activities	
5	Asset Turnover	Baranes and Palas (2019)[28]
6	Current Liabilities	Milosevic (2016)[23], Huang (2019)[29], Huang, Capretz and Ho (2021)[2]
7	Current Asset	
8	Total Asset	
9	Total Liabilities	
10	Earning Per Share	Milosevic (2016)[23], Prasad et al. (2022)[30]
11	D/E	
12	ROA	
13	Net Margin	Milosevic (2016)[23], Yang, Liu and Wu (2018)[31], Jones, Moser, & Wieland, (2020)[32]
14	Revenue	
15	Net Profit	
16	Current Ratio	
177	P/E	Milosevic (2016)[23], Torres P et al. (2019)[33], Baranes and Palas (2019)[28]
18	P/B	
19	Book Value Per Share	

Source: Author

and downward market trends[35]. Empirical research by Kim et al. (2020) supports the efficacy of SVC in financial forecasting[36]. They applied SVC to predict stock price movement, emphasizing its ability to handle nonlinear relationships in financial data and outperform other traditional methods.

Logistic Regression is a fundamental and interpretable algorithm used primarily for binary classification tasks. It models the probability of an event occurrence based on a set of input variables. In our research, Logistic Regression provides a baseline model for understanding the relationship between financial indicators and stock market trends[37]. Empirical research by Pahwa et al. (2017) demonstrates the application of Logistic Regression in predicting stock price movements[38]. Their study found that Logistic Regression can effectively capture the probabilities of stock price increases, providing valuable insights for investment decisions.

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve predictive accuracy and handle complex relationships in data[39]. In our research, Random Forest is applied to capture the intricate interactions among various financial indicators that may influence stock market trends. Its ability to reduce overfitting and provide robust predictions makes it a valuable tool for comprehensive financial analysis. Empirical research by Breiman (2001) [6] and Liaw and Wiener (2002)[40] highlights the effectiveness of Random Forest in handling large datasets and complex feature interactions.

RESULTS & DISCUSSION

Descriptive Analysis

As Figure 1, the analysis reveals that certain pairs of financial metrics, such as ”Cash” with ”Revenue,” ”Cash” with ”Current Asset,” and ”Cash” with ”Total Asset,” exhibit a high positive correlation, indicated by correlation values approaching 1. This suggests a proportional relationship, where an increase in one variable is likely mirrored by an increase in the other. Conversely, the variable pair ”Earnings Per Share (EPS)” and ”Price to Earnings (P/E) Ratio” display a negative correlation, specifically -0.110540. This indicates an inverse relationship, where an increase in EPS, which signifies higher earnings per share, tends to result in a decrease in the P/E ratio, suggesting the stock may be undervalued or earnings are on the rise relative to the share price.

Variables such as ”Net Profit” and ”Total Liabilities” present a correlation value near zero, denoting a lack of linear relationship between the two, suggesting that the net profit of a company is not directly affected by its total liabilities within the observed data set. Furthermore, the heatmap provides insights into the correlations between financial indicators like ”Cash,” ”Revenue,” ”Net Profit,” ”Current Asset,” ”Total Asset,” ”Total Liabilities,” ”Current Liabilities,” ”EPS,” and ”BVPS” (Book Value Per Share), along with the ”P/E” Ratio. These correlations reflect the interdependencies between these financial metrics. Lastly, the heatmap depicts the relationships between financial metrics and business operations, evident from the correlation values between ”Cash from Operating Activities, Cumulative,” ”Cash from Investing Activities, Cumulative,” and ”Cash from Financing Activities, Cumulative” with other financial indicators. These correlations are crucial as they highlight the impact of business activities on the financial health of the company.

Algorithm Performance Evaluation

Table 2 describes the comparison of machine learning algorithms for classification tasks, the performance metrics of three widely used classifiers were evaluated: Support Vector Classifier (SVC), Logistic Regression, and Random Forest. These classifiers were assessed based on accuracy, F1 score, precision, and recall—metrics that are pivotal in understanding the classifiers’ performance nuances. The Support Vector Classifier (SVC) achieved the highest overall accuracy at 59.1%, indicative of its robust generalization capabilities in classifying instances correctly. However, its F1 score of 0.572, while respectable, was not the highest observed, reflecting a potential compromise in the balance between precision and recall. Notably, the SVC’s precision of 0.551 was the lowest among the classifiers, suggesting a propensity to classify negative instances as positive.

Contrastingly, Logistic Regression, with an accuracy slightly trailing at 58.1%, demonstrated superior recall at 70.8%. This high recall rate underscores the model’s strength in identifying most positive instances, a desirable feature in domains were failing to detect positives is critically disadvantageous. However, this comes at the cost of precision, which at 0.533 is the lowest amongst the models, implying that while it captures most positives, it also incurs a higher rate of false positives. The Random Forest classifier presented a balanced but middling performance with nearly identical precision and recall scores (0.543 and 0.551, respectively). While its accuracy is comparable to that of Logistic Regression, its F1 score of 0.547 is the lowest, suggesting an overall weaker performance in terms of precision-recall balance.

From a comparative perspective, the choice of algorithm seems to be a function of the specific requirements of the classification task at hand. If overall accuracy is the criterion of paramount importance, the SVC emerges as the leading choice. Conversely, for applications where the cost of missing a positive is substantial, Logistic Regression would be preferred despite its lower precision. Random Forest, with its equitable precision and recall, could be considered when a balance between type I and type II errors is essential. These insights highlight the intrinsic trade-offs that practitioners must navigate when selecting a machine learning algorithm for predictive modeling in various domains of application.

For more detail, Table 3 shows the classification report for the Support Vector Classifier (SVC) shows moderately balanced performance metrics. Precision is higher for the ”Decrease/Unchanged (0)” category at 0.61, suggesting better accuracy in predicting non-increases, while the ”Increase (1)” category has a lower precision of 0.53, indicating more false positives in predictions of increases. The recall rates are similar for both categories, hovering around the mid-50s in percentage terms, which points to a moderate ability to identify true positives. The overall accuracy stands at 0.57, indicating that the model correctly predicts 57% of the outcomes. The Macro and Weighted Averages are identical at 0.57 across precision, recall, and F1-Score, showing that the model’s performance is consistent across classes, and there’s no significant bias introduced by class imbalance. The F1-Scores for both classes are also similar, suggesting a balanced trade-off between precision and recall, but

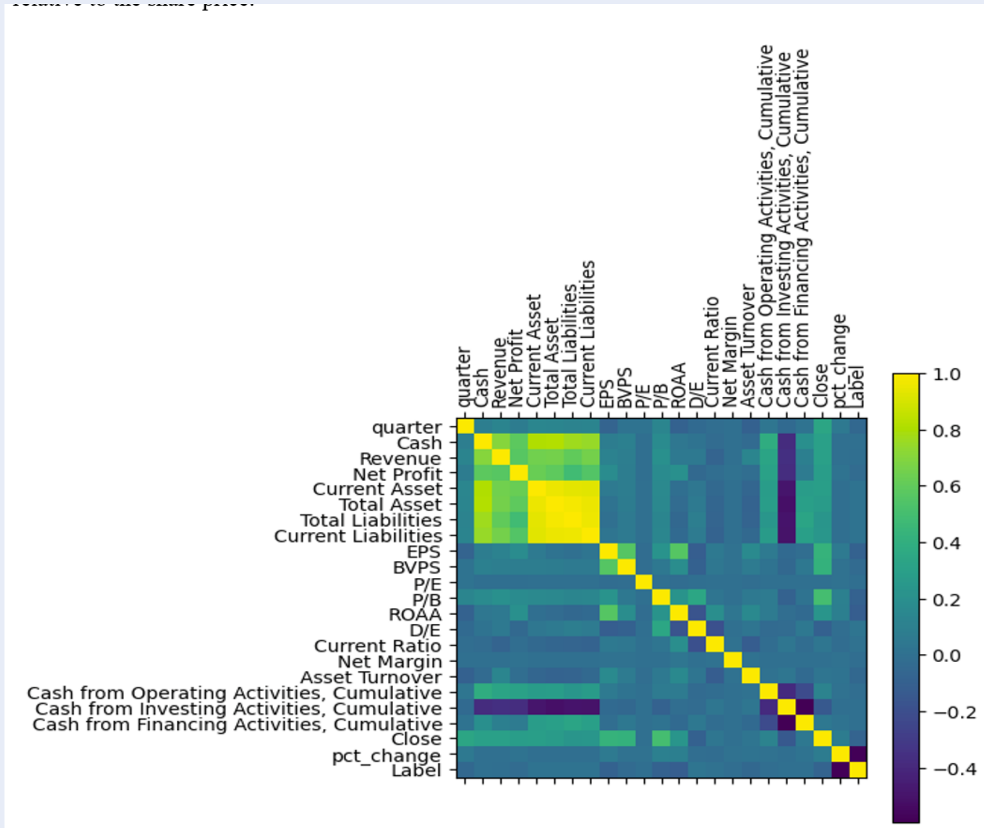

Figure 1: Variables correlation (Source: by author's calculation)

Table 2: Algorithm Performance Evaluation

Algorithm	accuracy	f1	precision	recall
SVC	0.591	0.572	0.551	0.594
Logistic Regression	0.581	0.609	0.533	0.708
Random Forest	0.580	0.547	0.543	0.551

Source: by author's calculations

they also indicate that there is room for improvement in the model's predictive accuracy.

Next, Table 4 states the classification report for Logistic Regression presents a nuanced performance when compared with the SVC. For the "Decrease/Unchanged (0)" category, Logistic Regression shows higher precision than the SVC (0.64 vs. 0.61) but a lower recall (0.50 vs. 0.57), indicating it is more selective but less sensitive in predicting non-increases. For the "Increase (1)" category, it has a similar precision to the SVC (0.54 vs. 0.53) but a notably higher recall (0.68 vs. 0.58), suggesting it is better at identifying true increases.

The overall accuracy for Logistic Regression is marginally higher at 0.59 compared to the SVC's 0.57. The Macro and Weighted Averages for Logistic Regression are slightly higher than those of the SVC, reflecting a small overall improvement in performance across the classes. The F1-Scores also show a similar pattern, with the score for "Increase (1)" being notably better in Logistic Regression (0.60 vs. 0.56), while the score for "Decrease/Unchanged (0)" is marginally lower (0.56 vs. 0.59). In summary, Logistic Regression appears to be more accurate and balanced overall compared to the SVC, with strengths in identifying increases.

Finally, Table 5 shows the classification report for Random Forest indicates an improvement over both SVC and Logistic Regression. It demonstrates higher

Table 3: Classification report for SVC

	Precision	Recall	F1-Score	Support
Decrease /unchanged (0)	0.61	0.57	0.59	1512
Increase (1)	0.53	0.58	0.56	1287
Accuracy			0.57	2799
Macro Avg	0.57	0.57	0.57	2799
Weighted Avg	0.58	0.57	0.57	2799

Source: by author's calculations

Table 4: Classification report for Logistic Regression

	Precision	Recall	F1-Score	Support
Decrease /unchanged (0)	0.64	0.5	0.56	1512
Increase (1)	0.54	0.68	0.6	1287
Accuracy			0.59	2799
Macro Avg	0.59	0.59	0.58	2799
Weighted Avg	0.6	0.59	0.58	2799

Source: by author's calculations

precision and recall for the "Decrease/Unchanged (0)" category compared to both previous models, with a precision of 0.64 (equal to Logistic Regression and higher than SVC's 0.61) and a recall of 0.61 (higher than both SVC's 0.57 and Logistic Regression's 0.50). For the "Increase (1)" category, Random Forest shows a slight improvement in precision over SVC and Logistic Regression and a comparable recall to SVC.

The overall accuracy of Random Forest is the highest at 0.60, slightly better than Logistic Regression's 0.59 and notably better than SVC's 0.57. Both the Macro and Weighted Averages for Random Forest are uniformly 0.60, indicating a consistent performance across the board and surpassing the averages for SVC and Logistic Regression. In essence, Random Forest outperforms the other two models in accuracy and maintains a balanced precision-recall across classes, showing it to be the most effective model among the three based on these metrics.

The empirical results indicate that all three models perform stably in constructing predictive outcomes for each category. With an accuracy of 60%, the highest among the three models, Random Forest remains the most effective and suitable algorithm for classification tasks. However, in the case of the author's dataset, the results are lower than the previous study where the predictive outcomes of the model were above 70%. The limitations within the scope of the study, the characteristics of the market in Vietnam, and a smaller scale might lead to these differences in results. When trading based on the Random Forest model, if investors avoid stocks that decrease or do not increase over the next three months combined with market research and the predicted increase in business operations, the outcomes remain promising for investors to trade based on their own risk tolerance and profit-seeking.

CONCLUSION & RECOMMENDATION

Conclusion

In conclusion, this research highlights the nuanced performance of machine learning algorithms in predicting stock price trends. The Random Forest algorithm emerged as the most effective, demonstrating a superior balance in precision and recall. This finding is particularly insightful given the complex nature of the Vietnamese stock market. The study's results, contrasting with the lower precision yet higher recall of Logistic Regression and the modest performance of Support Vector Classifier, underscore the importance of choosing the right algorithm based on specific market characteristics and data qualities.

The effectiveness of the Random Forest algorithm in your study, particularly for predicting stock price trends, lies in its ability to manage complex, nonlinear data typical of the stock market. Its balanced approach to classification helps navigate the intricacies of financial data, making it a robust tool for capturing the dynamic and often unpredictable movements in stock prices. This suitability for handling multifaceted financial datasets highlights Ran-

Table 5: Classification report for Random Forest

	Precision	Recall	F1-Score	Support
Decrease /unchanged (0)	0.64	0.61	0.62	1512
Increase (1)	0.56	0.59	0.58	1287
Accuracy			0.6	2799
Macro Avg	0.6	0.6	0.6	2799
Weighted Avg	0.6	0.6	0.6	2799

Source: by author's calculations

dom Forest as a highly applicable model for stock market analysis, especially in markets with intricate patterns and volatility like those in Vietnam.

Th conclusion of this study in Vietnam's stock market aligns with Patel et al. (2015)[12] and Vijh et al. (2020)[5], emphasizing machine learning's varied effectiveness based on data characteristics. However, your focus on a traditional model contrast with the trend towards deep learning in complex datasets highlighted by Shen and Shafiq (2020)[14] and Wang, Fan, and Wang (2021)[15]. Your research reinforces the relevance of context-specific model selection, showcasing Random Forest's robustness in markets like Vietnam, despite the global shift towards advanced deep learning models. Also, this conclusion intertwines the empirical findings with the theoretical backdrop of Behavioral Finance and the Efficient Market Hypothesis. It suggests that while advanced machine learning techniques like Random Forest, as shown in your research, offer robust predictions in certain market contexts like Vietnam, these tools also bring a new dimension to the classic debate between rational market behavior and behavioral influences. This highlights the ongoing evolution and complexity of financial markets, where both traditional models and emerging machine learning techniques are essential to capture the full spectrum of market dynamics.

Recommendation

Investors should recognize that machine learning, especially Random Forest, has the potential to revolutionize their stock market analysis in emerging markets such as Vietnam. By harnessing the power of these advanced algorithms, investors can unlock deeper insights into market trends, helping them make data-driven decisions that account for complex variables. Moreover, the synergy between diversification strategies and technological advancements in financial analysis can create a comprehensive investment approach that balances risk and returns while staying adaptable in an ever-evolving financial landscape. Furthermore, the dynamic nature of emerging markets demands adaptability, and machine learning models offer the flexibility to adjust investment strategies rapidly based on evolving market conditions. By integrating these models, investors can gain a competitive edge and navigate the intricate landscape of emerging markets with greater precision, potentially yielding more successful investment outcomes.

The findings suggest that managers in financial institutions should incorporate machine learning insights into their investment and risk assessment strategies. Understanding the strengths of different algorithms, like the effectiveness of Random Forest in specific market conditions, can aid in better portfolio management and decision-making processes. Regular training and updates on the latest financial technologies and machine learning applications could also be beneficial. With the exponential growth of financial data, machine learning provides a scalable and efficient way to analyze vast datasets, identify trends, and generate actionable insights. This can significantly improve the speed and accuracy of decision-making, enabling financial institutions to adapt swiftly to market shifts and customer preferences. Moreover, the ability of machine learning to detect subtle patterns and anomalies enhances risk assessment, allowing for more proactive risk mitigation strategies. In essence, the integration of machine learning is not just a technological advancement but a strategic imperative for financial managers looking to remain competitive and resilient in a rapidly evolving industry.

Financial institutions can harness these insights to gain a deeper understanding of market dynamics and enhance their risk assessment and investment strategies. The proficiency of machine learning models in forecasting market trends can lead to more informed decision-making within these institutions, particularly in the context of emerging and volatile markets. Furthermore, promoting an environment that encourages innovation in financial technologies while upholding market stability and investor protection is paramount. Regulators can play a pivotal role in achieving this balance by leveraging machine learning insights to inform their regulatory policies and surveillance mechanisms.

The research, while providing valuable insights, does have certain limitations that should be acknowledged. One notable limitation pertains to the scope of the data used. The analysis heavily relies on historical data, which might not fully capture the nuances of rapidly changing market conditions, especially in emerging markets. Moreover, the effectiveness of machine learning algorithms, such as Random Forest, can be influenced by the quality and completeness of the data available. Inaccurate or incomplete data may lead to suboptimal results. Additionally, the research primarily focuses on the application of specific algorithms and may not account for the evolving landscape of machine learning techniques. As the field of machine learning continues to advance, newer algorithms may outperform those discussed in this research. Therefore, the findings should be interpreted within the context of these data and algorithmic limitations to ensure a comprehensive understanding of the research's scope and implications.

This research, while providing valuable insights, does have certain limitations that should be acknowledged. One notable limitation pertains to the scope of the data used. The analysis heavily relies on historical data, which might not fully capture the nuances of rapidly changing market conditions, especially in emerging markets. To address this limitation, future research could incorporate real-time data sources and sentiment analysis to provide a more dynamic and up-to-date perspective on market trends. Moreover, the effectiveness of machine learning algorithms, such as Random Forest, can be influenced by the quality and completeness of the data available. Inaccurate or incomplete data may lead to suboptimal results. To mitigate this limitation, researchers can explore data cleansing and augmentation techniques, ensuring that the input data is as accurate and comprehensive as possible.

Additionally, the research primarily focuses on the application of specific algorithms and may not account for the evolving landscape of machine learning techniques. As the field of machine learning continues to advance, newer algorithms may outperform those discussed in this research. Future studies could explore a broader range of machine learning models and their applications in financial analysis to ensure a comprehensive understanding of the evolving landscape.

In conclusion, while acknowledging these limitations, this research provides a strong foundation for further exploration. Future research endeavors should aim to overcome these constraints, incorporating real-time data, improving data quality, and exploring a wider array of machine learning algorithms. This approach will not only enhance the robustness of financial analysis but also contribute to a more comprehensive understanding of the dynamic nature of the financial markets and the evolving role of machine learning within them.

FUNDING

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

ABBREVIATIONS

SVC: Support Vector Classification
SVM: Support Vector Machines
EMH: Efficient Market Hypothesis
ANN: Artificial Neural Networks
RMSE: Root-Mean-Square Deviation
MAPE: Mean Absolute Percentage Error
KNN: K-Nearest Neighbors
LSTM: Long Short-Term Memory
HOSE: The Ho Chi Minh Stock Exchange
IQR: The Interquartile Range
P/E: Price to Earnings
D/Y: Dividend Yield
EPS: Earning Per Share
K-S: Kolmogorov-Smirnov Test
TATO: Total Assets Turnover
B/M: Book to Market ratio
INTE: Interest Coverage Ratio
ROE: Return on Equity
F&B: Food and Beverage
ROA: Return on Asset
CR: Current Ratio
BVPS: Book Value Per Share

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Phan Huy Tam: Background theories, reviewing and providing feedbacks on the manuscript.
Doan Thi Ngoc Dieu: Analyzing data, Abstract, Introduction, Data and Methodology, Result and Discussion, Conclusion and Recommendations, References.

REFERENCES

1. Agrawal M, Shukla PK, Nair R, Nayyar A, Masud M. Stock Prediction Based on Technical Indicators Using Deep Learning Model. Comput Mater Continua. 2022;70(1);Available from: https://doi.org/10.32604/cmc.2022.014637.
2. Huang Y, Capretz LF, Ho D. Machine learning for stock prediction based on fundamental analysis. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI); 2021;Available from: https://doi.org/10.1109/SSCI50451.2021.9660134.
3. Mokhtari S, Yen KK, Liu J. Effectiveness of artificial intelligence in stock market prediction based on machine learning. arXiv preprint arXiv:2107.01031; 2021;Available from: https://doi.org/10.5120/ijca2021921347.
4. Nabipour M, Nayyeri P, Jabani H, Shahab S, Mosavi A. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access. 2020;8:150199-212;Available from: https://doi.org/10.1109/ACCESS.2020.3015966.
5. Vijh M, Chandola D, Tikkiwal VA, Kumar A. Stock closing price prediction using machine learning techniques. Procedia Comput Sci. 2020;167:599-606;Available from: https://doi.org/10.1016/j.procs.2020.03.326.
6. Soni P, Tewari Y, Krishnan D. Machine Learning approaches in stock price prediction: A systematic review. J Phys Conf Ser. 2022;Available from: https://doi.org/10.1088/1742-6596/2161/1/012065.
7. Raei S, Fallahpour S. Behavioral finance. A different approach in financial arena. Tehran Univ Financ Res Q. 2004;18:77-106;.
8. Fama EF. Efficient capital markets: A review of theory and empirical work. J Finance. 1970;25(2):383-417;Available from: https://doi.org/10.2307/2325486.
9. Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertainty. 1992;5:297-323;Available from: https://doi.org/10.1007/BF00122574.
10. Baker M, Wurgler J. Investor sentiment in the stock market. J Econ Perspect. 2007;21(2):129-51;Available from: https://doi.org/10.1257/jep.21.2.129.
11. Bekaert G, Harvey CR. Emerging equity market volatility. J Financ Econ. 1997;43(1):29-77;Available from: https://doi.org/10.1016/S0304-405X(96)00889-6.
12. Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Syst Appl. 2015;42(1):259-68;Available from: https://doi.org/10.1016/j.eswa.2014.07.040.
13. Nabipour M, Nayyeri P, Jabani H, Mosavi A, Salwana E. Deep learning for stock market prediction. Entropy. 2020;22(8):840;PMID: 33286613. Available from: https://doi.org/10.3390/e22080840.
14. Shen J, Shafiq MO. Short-term stock market price trend prediction using a comprehensive deep learning system. J Big Data. 2020;7(1):1-33;PMID: 32923309. Available from: https://doi.org/10.1186/s40537-020-00333-6.
15. Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognit Lett. 2021;141:61-67;Available from: https://doi.org/10.1016/j.patrec.2020.07.042.
16. Ngoc Hai P, Manh Tien N, Trung Hieu H, Quoc Chung P, Thanh Son N, Ngoc Ha P, et al. An Empirical Research on the Effectiveness of Different LSTM Architectures on Vietnamese Stock Market. In: Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System; 2020;Available from: https://doi.org/10.1145/3437802.3437827.
17. Khoa BT, Huynh TT. Forecasting stock price movement direction by machine learning algorithm. Int J Electr Comput Eng. 2022;12(6):6625;Available from: https://doi.org/10.11591/ijece.v12i6.pp6625-6634.
18. Khiem NM, Takahashi Y, Dong KTP, Yasuma H, Kimura N. Predicting the price of Vietnamese shrimp products exported to the US market using machine learning. Fish Sci. 2021;87:411-23;Available from: https://doi.org/10.1007/s12562-021-01498-6.
19. Huynh HD, Dang LM, Duong D. A new model for stock price movements prediction using deep neural network. In: Proceedings of the 8th International Symposium on Information and Communication Technology; 2017;Available from: https://doi.org/10.1145/3155133.3155202.
20. Le Hong H, Nguyen NN, Nguyen TL, Nguyen LD, Nguyen NH. Stock Market Prediction: The Application of Text-Mining in Vietnam. VNU J Econ Bus. 2022;2(2);.
21. Naknok S. Firm Performance Indicators as a Fundamental Analysis of Stocks and a Determinant of a Firm's Operation. Int J Econ Bus Adm (IJEBA). 2022;10(1):190-213;Available from: https://doi.org/10.35808/ijeba/758.
22. Dimitrantzou C, Psomas E, Vouzas F. The influence of competitive strategy and organizational structure on the cost of quality in food and beverage (F&B) companies. TQM J. 2023;Available from: https://doi.org/10.1108/TQM-01-2023-0031.
23. Milosevic N. Equity forecast: Predicting long term stock price movement using machine learning. arXiv preprint arXiv:1603.00751; 2016;.
24. Amel-Zadeh A, Calliess JP, Kaiser D, Roberts S. Machine learning-based financial statement analysis [Internet]. SSRN; 2020. Available from: SSRN 3520684;Available from: https://doi.org/10.2139/ssrn.3520684.
25. Rouf N, Malik MB, Arif T, Sharma S, Singh S, Aich S, et al. Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. Electronics. 2021;10(21):2717;Available from: https://doi.org/10.3390/electronics10212717.
26. Hu Z, Zhao Y, Khushi M. A survey of forex and stock price prediction using deep learning. Appl Syst Innov. 2021;4(1):9;Available from: https://doi.org/10.3390/asi4010009.
27. Kotios D, Makridis G, Fatouros G, Kyriazis D. Deep learning enhancing banking services: a hybrid transaction classification and cash flow prediction approach. J Big Data. 2022;9(1):100;PMID: 36213092. Available from: https://doi.org/10.1186/s40537-022-00651-x.
28. Baranes A, Palas R. Earning movement prediction using machine learning-support vector machines (SVM). J Manag Inf Decis Sci. 2019;22(2):36-53;.
29. Huang Y. Machine learning for stock prediction based on fundamental analysis [dissertation]. The University of Western Ontario (Canada); 2019;.
30. Prasad VV, Gumparthi S, Venkataramana LY, Srinethe S, Sruthi Sree R, Nishanthi K. Prediction of stock prices using statistical and machine learning models: a comparative analysis. Comput J. 2022;65(5):1338-51;Available from: https://doi.org/10.1093/comjnl/bxab008.
31. Yang H, Liu XY, Wu Q. A practical machine learning approach for dynamic stock recommendation. In: 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE); 2018;Available from: https://doi.org/10.1109/TrustCom/BigDataSE.2018.00253.
32. Jones S, Moser WJ, Wieland MM. Machine learning and the prediction of changes in profitability. Contemp Account Res. 2020;.
33. Torres EP, Hernández-Álvarez M, Torres Hernández EA, Yoo SG. Stock market data prediction using machine learning techniques. In: Information Technology and Systems: Proceedings of ICITS 2019; 2019;Available from: https://doi.org/10.1007/978-3-030-11890-7_52.
34. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273-97;Available from: https://doi.org/10.1007/BF00994018.
35. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999;9:293-300;Available

from: https://doi.org/10.1023/A:1018628609742.
36. Kim S, Ku S, Chang W, Song JW. Predicting the direction of US stock prices using effective transfer entropy and machine learning techniques. IEEE Access. 2020;8:111660-82;Available from: https://doi.org/10.1109/ACCESS.2020.3002174.
37. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. Vol. 398. John Wiley & Sons; 2013;Available from: https://doi.org/10.1002/9781118548387.
38. Pahwa N, Khalfay N, Soni V, Vora D. Stock prediction using machine learning a review paper. Int J Comput Appl. 2017;163(5):36-43;Available from: https://doi.org/10.5120/ijca2017913453.
39. Breiman L. Random forests. Mach Learn. 2001;45:5-32;Available from: https://doi.org/10.1023/A:1010933404324.
40. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18-22;.

Tạp chí Phát triển Khoa học và Công nghệ – Economics-Law and Management 2024, ():1-1

Open Access Full Text Article

Bài nghiên cứu

Ứng dụng máy học trong dự đoán xu hướng giá chứng khoán của các doanh nghiệp niêm yết trên Sở giao dịch chứng khoán Thành phố Hồ Chí Minh

Phan Huy Tâm[1,2,*], Đoàn Thị Ngọc Diệu[1,2]

TÓM TẮT

Nghiên cứu này đánh giá hiệu suất của các thuật toán máy học trong việc dự báo xu hướng giá cổ phiếu của các doanh nghiệp niêm yết trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh, tập trung vào các ngành ngoài ngân hàng, bảo hiểm và chứng khoán. Dữ liệu nghiên cứu từ năm 2015 đến năm 2022, bao gồm dữ liệu lịch sử giá cổ phiếu và các chỉ số tài chính cơ bản của doanh nghiệp. Bằng cách áp dụng các thuật toán máy học phổ biến như Phân loại Vector Hỗ trợ, Hồi quy Logistic và Rừng Ngẫu nhiên, nghiên cứu này đánh giá và xem xét thuật toán hiệu quả nhất cho dự đoán xu hướng chính xác. Kết quả cho thấy thuật toán Rừng Ngẫu nhiên vượt trội hơn các thuật toán khác, mang lại một cách tiếp cận cân bằng giữa độ chính xác và tỷ lệ bỏ sót. Phát hiện này rất hữu ích đối với các nhà đầu tư và các nhà phân tích tài chính trong việc đưa ra quyết định tài chính và đầu tư, đặc biệt là trong bối cảnh một thị trường đang phát triển và năng động như Việt Nam. Nghiên cứu nhấn mạnh sức mạnh của máy học trong dự báo tài chính, làm nổi bật tiềm năng của nó trong việc cách mạng hóa chiến lược đầu tư. Kết luận của nghiên cứu làm rõ tầm quan trọng của việc tích hợp các công cụ máy học, đặc biệt là Rừng Ngẫu nhiên, trong phân tích tài chính và quá trình ra quyết định. Nghiên cứu này không chỉ cung cấp một công cụ thực tế cho các nhà đầu tư mà còn đóng góp đáng kể vào tài liệu học thuật về dự đoán thị trường tài chính bằng các phương pháp luận máy học.

Từ khoá: dự báo xu hướng cổ phiếu, máy học, chỉ số tài chính

[1]*Trường Đại học Kinh tế - Luật, TP Hồ Chí Minh, Việt Nam.*

[2]*Đại học Quốc gia Thành phố Hồ Chí Minh, TP Hồ Chí Minh, Việt Nam.*

Liên hệ

Phan Huy Tâm, Trường Đại học Kinh tế - Luật, TP Hồ Chí Minh, Việt Nam.

Đại học Quốc gia Thành phố Hồ Chí Minh, TP Hồ Chí Minh, Việt Nam.

Email: tamph@uel.edu.vn

Lịch sử

- Ngày nhận: 17-1-2024
- Ngày sửa đổi: 30-5-2024
- Ngày chấp nhận: 15-8-2024
- Ngày đăng:

DOI :

Trích dẫn bài báo này: Tâm P H, Diệu D T N. **Ứng dụng máy học trong dự đoán xu hướng giá chứng khoán của các doanh nghiệp niêm yết trên Sở giao dịch chứng khoán Thành phố Hồ Chí Minh**. *Sci. Tech. Dev. J. - Eco. Law Manag.* 2024; ():1-1.