Article

On the Predictability of Greek Systemic Bank Stocks using Machine Learning Techniques

Hera Antonopoulou[1, *], Leonidas Theodorakopoulos[2], Constantinos Halkiopoulos[3] and Vicky Mamalougkou[4]

[1] Department of Management Science and Technology, University of Patras, Greece; hera@upatras.gr
[2] Department of Management Science and Technology, University of Patras, Greece; theodleo@upatras.gr
[3] Department of Management Science and Technology, University of Patras, Greece; halkion@upatras.gr
[4] Department of Management Science and Technology, University of Patras, Greece; up1081371@upatras.gr
* Correspondence: hera@upatras.gr

Abstract: Background/Objectives: Accurate prediction of stock prices is an extremely challenging task because of factors such as political conditions, global economy, unexpected events, market anomalies, and relevant companies' features. In this work, the random forest has been used to forecast the prices of the four major Greek systemic banks Methods/Analysis: We make use of a set of financial variables based on intraday data: (i) Open stock price, (ii) High stock price, (iii) Low stock price, and (iv) Close stock price of a particular Greek systemic bank. Results/Findings: The variables used here are crucial in predicting systemic banks' stock closing prices. These provide a better prediction of the next day's closing price of the bank series. Novelty /Improvement: To our knowledge, this is the first study that employs machine learning techniques in Greek systemic banks.

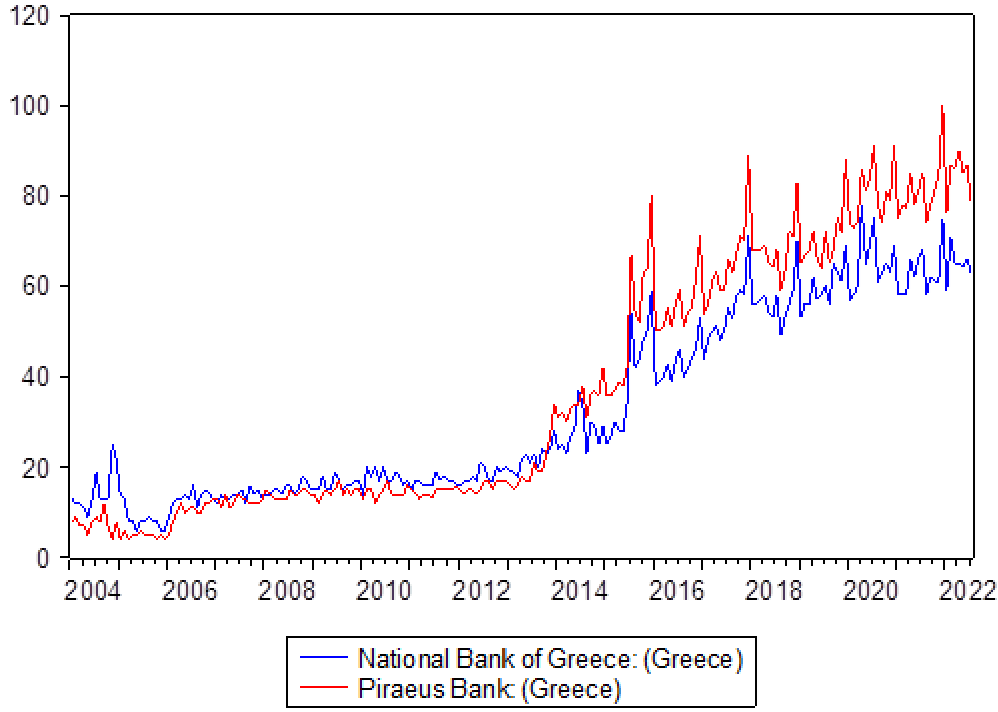
Keywords: Machine Learning; Random Forest; Google Trends; Predictability; Banks; Greece.

1. Introduction

Financial markets are known for their volatile, dynamic, non-linear nature [1]. Accurate prediction of stock prices is an extremely challenging task because of factors such as political conditions, global economy, unexpected events, market anomalies, relevant companies' features, etc. [2]. Considering that financial markets are linked to a country's economic conditions attracting major funds from investors and issues equities in the public interest, it is, therefore, crucial to forecast the movement of the stock prices to prevent excessive losses and make relevant investment decisions [3]. Traditionally, two main approaches have been widely employed in stock price predictions [4]. These approaches include technical analysis methods that utilize historical stock prices to predict future prices [5]. Further, the analysis is based on external factors considering information regarding Google Tends data based on overall market conditions and macroeconomic issues. The Google Trends Service is a free, cross-platform service from Google that analyzes a huge volume of searches and provides data on user search terms based on geographic and temporal patterns. That is, data about the volume of the user's search terms (search volume).

In this work, random forest (RF hereafter) has been used to forecast the prices of the four major Greek banks. We focus on Greek systemic banks since the recent years there is an increasing interest in the performance of the Greek and European banking sector. As we can see form the Figure 1 and Google Trends indices for the National Bank and Piraeus Bank, the online interest dramatically increased. The use of Google Trends indices has been widely explored in the related literature concluding that Google Trends improve forecasts' accuracy. Google Trends gives access to the number of actual search requests in the Google search engine, allowing us to measure people's interest in a particular topic for several languages and regions worldwide. From a financial point of view, we make use of a set of financial variables based on intraday data with (i) Open stock price, (ii) High stock price, (iii) Low stock price, and (iv) Close stock price of a particular Greek systemic bank. These data are expected to be crucial in order to better predict the next day's closing price of the bank series. Further, the accuracy model prediction is evaluating the most known performance measures in the existing literature: (i) RMSE and (ii) MAPE.

Figure 1. This is a figure. Schemes follow the same formatting.



2. Literature Review

Recently, more sophisticated machine learning methods are used to forecast stock prices, that follow either technical or fundamental analysis, considering the nature of the data in terms of data size and nonlinearities in the dataset [7]. In order to account for such data characteristics, more efficient models should be considered, that can effectively identify complex relations and hidden patterns in such large data sets [8]. Earlier studies in this area, mainly explore classic algorithms such as linear and non-linear regressions, random walk methods, moving average convergence/divergence, and several traditional linear econometrics models such as Autoregressive Integrated Moving Average (ARIMA) family models, to predict stock prices [9]. However, unlike conventional economic models, the use of machine learning can be more effective in predicting stock prices by analyzing the data in a more data-driven way. To this end, popular machine learning techniques found in the literature include SVM, Neural Networks, KNN, and Random Forests.

In their work, [10] explore the sensitivity of stock prices to external conditions including daily quotes of commodity prices such as gold, crude oil, natural gas, corn, and cotton with the results indicating that logistic regression performed better. [11] analyzed stock data containing daily stock information ranging from 2008 to 2013. Various algorithms were explored in a one-step ahead and multi-step ahead forecasting exercise of stock prices. Results revealed that the SVM machine learning technique produced better results in the long-term considering the highest accuracy score among the three models explored (Logistic Regression, Quadratic Discriminant Analysis, and SVM). [12] uses various technical indicators such as RSI, on balance Volume, and Williams %R among others, as features in the extremely randomized tree algorithm implemented. Most relevant features were then selected, that were further used as inputs to a Kernelized SVM model. In their work, [13] propose a hybrid model which utilizes cuckoo search in order to optimize the parameters of SVM. The given model explored several technical indicators such as RSI, Money Flow Index, EMA, Stochastic Oscillator, and MACD. [14] proposed the use of a neural network ensemble, to predict the direction of the stock price. [15] also used ensemble learning algorithms and specifically Random Forest to predict the direction of stock prices with technical indicators such as Relative Strength Index (RSI), stochastic oscillator used as inputs to train the model. The results indicate that the Random Forest algorithm outperforms existing algorithms explored in their study. [16], apply five machine learning techniques i.e., (i) Support Vector Machine, (ii) Random Forest, (iii) K-Nearest Neighbor (KNN), (iv) Naive Bayes, and finally, (v) SoftMax to predict stock market trends with the experimental results showing that Random Forest algorithm performs best for large datasets while Naïve Bayesian Classifier performs best for small datasets. [6] analyze several machine learning classification models concluding that the effectiveness and accuracy of the algorithm mainly depend on the type and volume of data on which predictions are analyzed.

[17] explore distinct machine learning methods (including support vector machine, random forest, and boosted decision trees, among others) in order to build prediction models and forecast the prices of the stocks for various exchange markets. [18], follow a combination of Technical and Fundamental analysis and further apply machine learning models i.e., Random Forest, Support Vector Machines, and a feed-forward Neural Network to predict the market using time series prediction and sentiment analysis. [19] focus on the Chinese market and propose comprehensive customization of feature engineering and deep learning-based model in order to predict the price trend of stocks in the examined market. The proposed system achieves high accuracy in terms of stock market trend prediction.

[20] compare different machine learning classification algorithms against the National Association of Securities Dealers Automated Quotations System (NASDAQ), New York Stock Exchange (NYSE), Nikkei, and Financial Times Stock Exchange (FTSE) concluding that Random Forest and Bagging with leaked dataset provides satisfactory performance. Additionally, [21], uses a random forest model to predict stock price movements. [22] focuses on predicting the Nifty 50 Index by examining eight Supervised Machine Learning Models (Random Forest, Linear regression, SVM, and kNN, among others). The evaluation results showed that SVM performed better than the other machine algorithms explored, but with an increase in the dataset size, Stochastic Gradient Descent gave better results. [23] employs Random Forest and Logistic regression on a five-year-long dataset of volume and price of the Tesla company stocks traded on the New York Stock Exchange (NYSE) to estimate the closing prices of these stocks.

Most of the methods above can be used for stock market forecasting prediction but may be employed broadly in other economics of financial series and more recently in the cryptocurrency market. Examples include among many others, [24]. Random Forest (RF) machine learning technique has been further employed by [25] and [26] to predict bitcoin volatility. In their work [27] employ Generalized Random Forests to estimate and predict the risk measure Value at Risk for Cryptocurrencies.

In line with the above-mentioned research in this field, we propose the use of Random Forest for forecasting purposes of stock prices. Random Forest is an ensemble technique, basically operating by building several decision trees at training time [28]. Predictions using Random Forests are generated by averaging the predictions of the decision trees which allows us to reduce the variance and improve the efficiency of test set Decision Trees and avoid overfitting.

The remainder of the papers is as follows: Section 3 gives the methods applied. Section 4 discusses the findings, and finally, section 5 concludes the paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2022 doi:10.20944/preprints202207.0462.v1

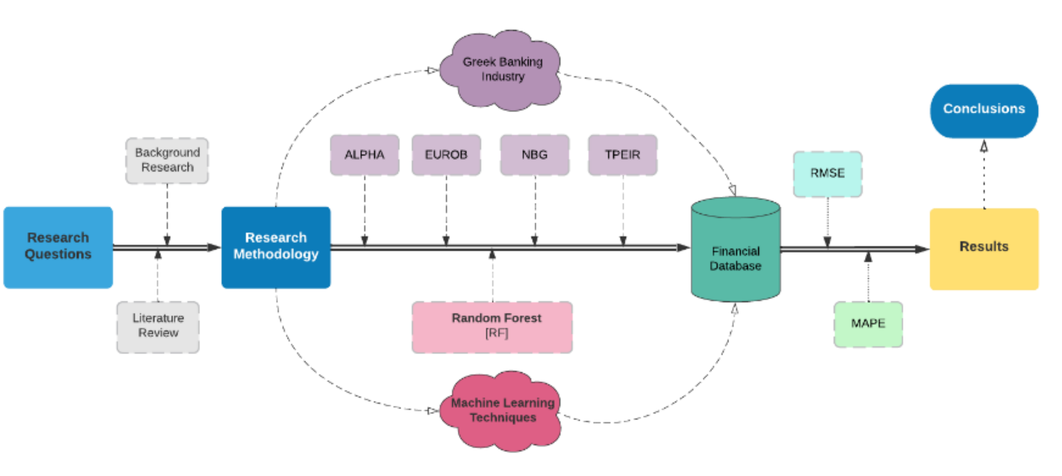


Figure 2. **This is a figure. Schemes follow the same formatting.**

3. Methodology

3.1. Description of Data

The data used here for the four Greek systemic banks have been collected from 2001. The dataset includes 20 years of data from 1/1/2001 to 30/12/2020 of Alpha Bank (ALPHA), National bank of Greece (NBG), Eurobank (EUROB), and Piraeus Bank (TREIR) incorporating various market phases such as booms and crashes. The data used here include intraday data information (i.e., (i) Open stock price, (ii) Close stock price, and (iii) Volume. Only the day-wise closing systemic bank stock price has been obtained. The rolling window methods have been used for model validation.

Once the dataset is available, it requires the application of some widely used pre-processing methods in order to obtain a better fitting to the model used [29]. The importance of the estimation context is in accordance with the pre-processing method of the dataset. Previous methods have shown that we should take into account the following major pre-processing stages: (i) feature selection, (ii) order reduction, (iii) and representation of features.

3.2. New Variables

Additional variables are used for the forecast of each of the systemic bank's closing prices. These variables have been employed for model training. The new variables considered are the following:

1. A new variable based on the bank closing price using a five-day moving average (noted by MA-5)
2. A new variable based on the bank closing price using a ten-day moving average (noted by MA-10)
3. A new variable based on the bank closing price using a twenty-one-day moving average (noted by MA-21)
4. A new variable based on the bank closing price using standard deviation for the past five days (noted by STD DEV-5)

3.3. Machine Learning Methods

This sub-section endeavors to sum up the machine learning techniques deployed in preceding research for stock prediction and forecasting see [30]. After the data is pre-processed and transformed into standard representation, they were placed in machine learning models for more extended processing. The following section summarily the various machine learning approaches presented: (i) Artificial Neural Networks, (ii) Support Vector Machine, (iii) Naïve Bayes, and (iv) Deep Neural Learning.

Artificial Neural Networks (ANN). A neural network stands for a network that is based on a simple computer node (neurons) interconnected. It is inspired by the Central Nervous System (CNS), which in fact it attempts to simulate. Their architecture is based on the architecture of Biological Neural Networks and so they are able to perform calculations in a massive way. The purpose of ANNs is to be able to perform the calculations performed by the human brain, that is, to be able to transmit information about the stimuli they receive. ANNs are trained to be able to solve the tasks assigned to them or to be able to perform certain processes on their own e.g., recognize images. First, however, it is essential that they are properly trained.

Support Vector Machine (SVM). SVMs stand for a set of supervised learning models with algorithms used for classification and regression analysis. The fundamental principles of SVM, developed by Vapnik, are based on the theory of statistical learning. They have been used in a wide range of real-world applications such as text categorization, image recognition, audio, data sorting, and data detection.

Naïve Bayes (NB). The Naive Bayes algorithm is a family of probabilistic machine learning models used for classification problems and their central idea is based on Bayes's theorem.

Deep Neural Learning (DNN). Deep Neural learning is a new approach to data learning. Uses a specific family of models: sequences of simple functions interconnected. These function chains are the neural networks. These sequences of functions can analyze a complex idea in a hierarchy of simpler ones. Each layer organizes the previous layer into more advanced and abstract concepts. Below we present the Random Forest procedure which was used for the empirical part of the paper as it gives the best results across the previous method presented.

3.4. Machine Learning Methods – Random Forest

Numerous machine learning and deep learning forecasting algorithms have been developed in recent years to address the increasing diversity and complexity of forecasting challenges. The selection of a machine learning algorithm is contingent on numerous variables, including the business question you are attempting to answer, the availability and relevance of historical data, the accuracy and success metric you must achieve, the horizon, and the amount of time your team must develop a forecasting solution [31]. These limits must be regularly and on multiple levels weighed. In the case of demand forecasting, load data must be continuously and regularly projected, and a robust and dependable data intake procedure must be in place to assure the flow of raw data [32].

Random Forest (RF) is an ensemble machine learning technique widely used for both classification and regression problems. Random forests are comprised of many individual arbitrary generated decision trees [28]. In regression, the Random Forest algorithm is basically described by the following: consider a regression tree which is consisted of branches, and internal and terminal nodes. By constructing the regression tree, at each step, we select features in a random way that comes from the set of predictor variables, and the best split- based on a pre-chosen cost function- is used to split the node. This random selection of predictor variables helps mitigate the influence of influential predictors on each Tree building. The building of the decision tree continues until the set of terminal nodes for the tree stopping criterion or number of observations on every terminal node is met. The final prediction using Random Forests is produced by averaging the predictions of the individual ensemble trees. This allows us to reduce the variance and improve the efficiency of test set Decision Trees and therefore, avoid overfitting and deliver higher accuracy predictions [28] In this study, newly created technical analysis tools are provided as variables for training.

4. Results

In order to assess the forecasting performance of the models used, on four different sector banks namely, Alpha Bank, Eurobank, National Bank of Greece, and Piraeus Bank with the use of RF model methods, we employ two widely used evaluation criteria: (i) the Root Mean Square Error (RMSE), (ii) and the Mean Absolute Percentage Error (MAPE).

The RMSE metric which has been utilized to evaluate the performance of the model is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)} \qquad (1)$$

where 'y_i' is the actual closing price series for each bank considered, '$\hat{y}_i$' is the estimation closing price value, while the symbol 'n' stands for the total size of the window considered.

The criterion MAPE is also employed for model evaluation performance. This criterion is given by the following:

$$MAPE = \frac{100}{n}\sum_{i}^{n}\frac{y_i - \hat{y}_i}{y_i} \qquad (2)$$

where 'y_i' again is the actual closing price series for each bank considered, '$\hat{y}_i$' is the estimation closing price value, while the symbol 'n' is the total window size considered.

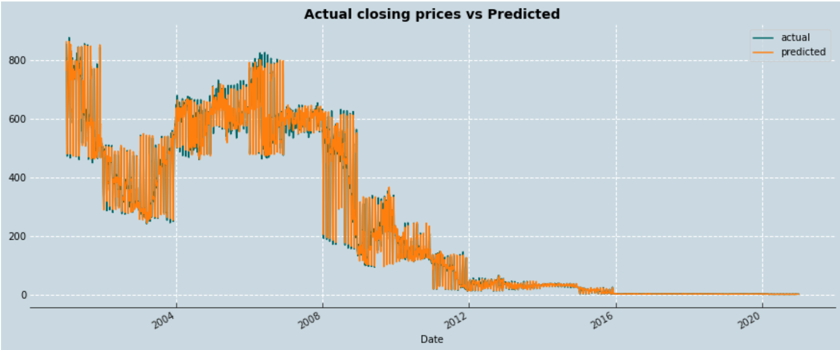


Figure 3: Alpha Bank closing price predicted vs actual values

Each of the following figures depicts the actual price of each of the four systemic Greek banks considered in regard to the forecast value obtain for the RF model employed. Further results for the RMSE and MAPE evaluation criteria are given in the following table (Table 1). In particular, the figure above (Figure 3) and below figures (Figure 4,5,6) shows a visual representation of the performance of the proposed machine learning models for each of the four banks considered.

We observe in all four cases that the random forest-based forecasting models predicted values follow for the most part similar patterns as the observed values.

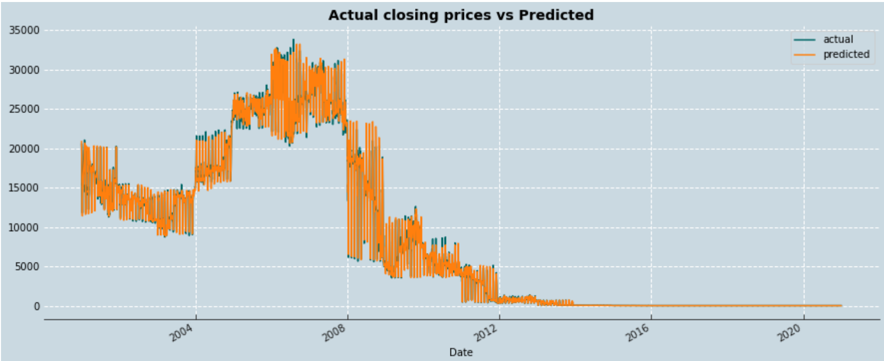


Figure 4: Eurobank closing price predicted vs actual values

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2022 doi:10.20944/preprints202207.0462.v1

Additionally, we observe that in most cases the predictions are very close to the actual values. Moreover, as it is evident from the plots in each case, the model has captured well upward or downward moves in the corresponding bank return series.

Furthermore, according to the results presented in Table 1 when considering the MAPE values obtained, among the four banks examined, the proposed models based on a one-step-ahead out-of-sample forecasting approach, performed better when predicting Alpha Bank returns.

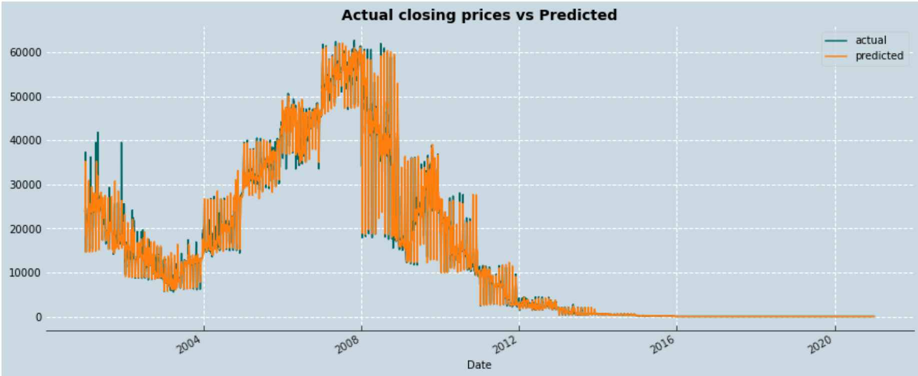


Figure 5: National Bank of Greece closing price predicted vs actual values

Moreover, we notice low magnitude differences between the MAPE values of three of the four cases (Alpha Bank, Eurobank, and Piraeus Bank). For the case of the National Bank of Greece, the corresponding MAPE value is 10.60%. Overall, we conclude that our proposed forecasting approach is able to accurately predict Greek bank sector returns.

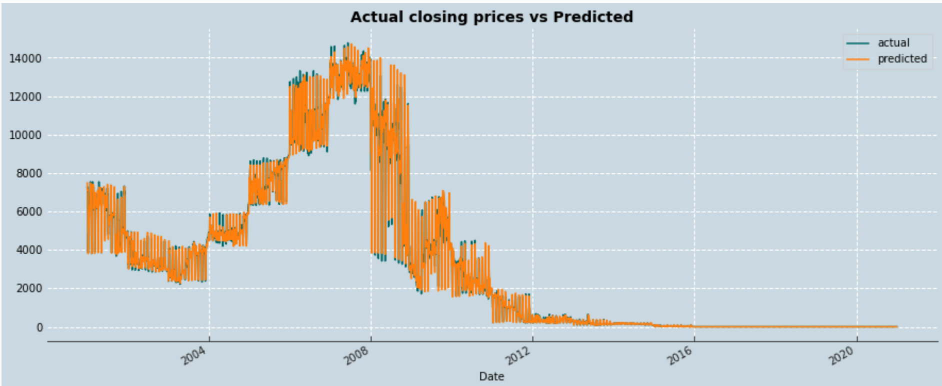


Figure 6: Piraeus Bank closing price predicted vs actual values

	RMSE	MAPE

Table 1. RMSE and MAPE values obtained using RF models

Alpha Bank	12.9450	4.7247
Eurobank	414.4279	5.1782
National Bank of Greece	1201.3288	10.6036
Piraeus Bank	185.7749	5.4684

5. Conclusions

Accurate forecast of bank prices is a challenging task capturing the interest of academics and investors. Several machine learning algorithms have been widely employed considering their effectiveness to identify complex relations in the stock market data. In this work, the Random Forest machine learning algorithm has been employed to predict stock prices in a one-step-ahead out-of-sample forecasting exercise. Additional technical analysis tools have been used as inputs to the machine learning model for each of the four cases of Greek banks consider in the analysis. Random Forest has been effectively used in this case also. We show that in all four cases the random forest-based forecasting models predicted values follow for the most part similar patterns as the observed values. We also observe that in most cases the predictions are very close to the actual values. Overall, we conclude that our proposed forecasting approach is able to accurately predict Greek bank sector returns. However, further research should be conducted considering more sophisticated machine learning and deep learning models. Future research could use Google trends data related to financial news variables to improve the effectiveness in predicting Greek banks' stock prices.

Author Contributions: H.A., L. T., C.H., and V.M., contributed to the design and implementation of the research, the analysis of the results, and the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Masoud, Najeb MH. (2017). "The impact of stock market performance upon economic growth." International Journal of Economics and Financial Issues 3 (4): 788–798.
2. Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the Chinese stock market. Heliyon, 5(8), e02310. https://doi.org/10.1016/j.heliyon.2019.e02310
3. Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine Learning Approaches in Stock Price Prediction: A Systematic Review. Journal of Physics: Conference Series, 2161(1). https://doi.org/10.1088/1742-6596/2161/1/012065
4. Murkute, Amod, and Tanuja Sarode. (2015). "Forecasting market price of stock using artificial neural network." International Journal of Computer Applications 124 (12): 11-15.
5. Hur, Jung, Manoj Raj, and Yohanes E. Riyanto. (2006). "Finance and trade: A cross-country empirical analysis on the impact of financial development and asset tangibility on international trade." World Development 34 (10): 1728-1741.
6. Misra, M., Yadav, A. P., & Kaur, H. (2018). Stock Market Prediction using Machine Learning Algorithms: A Classification Study. 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 2475–2478.
7. Li, Lei, Yabin Wu, Yihang Ou, Qi Li, Yanquan Zhou, and Daoxin Chen. (2017). "Research on machine learning algorithms and feature extraction for time series." IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC): 15.
8. Seber, George AF and Lee, Alan J. (2012). "Linear regression analysis." John Wiley &Sons 329
9. Yuqing Dai, Yuning Zhang (2013). Machine Learning in Stock Price Trend Forecasting. Stanford University.

10. G.S. Atsalakis, I.G. Atsalaki, F. Pasiouras, C. Zopounidis. (2019). Bitcoin price forecasting with neuro-fuzzy techniques, Eur. J. Oper. Res. 276 770–780. https://doi.org/10.1016/j.ejor.2019.01.040.
11. Zhang, G. Peter. (2003). "Time series forecasting using a hybrid ARIMA and neural network mode." Neurocomputing 50: 159-175.
12. Di, X. (2014). Stock Trend Prediction with Technical Indicators using SVM. Stanford University. Retrieved from http://finance.yahoo.com
13. Devi, K. N., Bhaskaran, V. M., & Kumar, G. P. (2015). Cuckoo optimized SVM for stock market prediction. In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1–5). https://doi.org/10.1109/ICIIECS.2015.7192906
14. Liaw, Andy, and Matthew Wiener. (2002) "Classification and regression by Random Forest." R news 2 (3): 18-22.
15. Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. 00(00), 1–20. http://arxiv.org/abs/1605.00003
16. Kumar, Manish, and M. Thenmozhi. (2006) "Forecasting stock index movement: A comparison of support vector machines and random forest" In Indian institute of capital markets 9th capital markets conference paper
17. Singh, S., Madan, T. K., Kumar, J., & Singh, A. K. (2019). Stock Market Forecasting using Machine Learning: Today and Tomorrow. 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 1, 738–745. https://doi.org/10.1109/ICICICT46008.2019.8993160
18. Pabuçcu, H., Ongan, S., Ongan. A. (2020). Forecasting the movements of Bitcoin prices: an application of machine learning algorithms, Quant. Financ. Econ. 4 679–692. https://doi.org/10.3934/qfe.2020031.
19. Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. Journal of Big Data, 7(1), 66. https://doi.org/10.1186/s40537-020-00333-6
20. Subasi, A., Amir, F., Bagedo, K., Shams, A., & Sarirete, A. (2021). Stock Market Prediction Using Machine Learning. Procedia Computer Science, 194, 173–179. https://doi.org/https://doi.org/10.1016/j.procs.2021.10.071
21. Srinu Vasarao, P., & Chakkaravarthy, M. (2022). Time Series Analysis Using Random Forest for Predicting Stock Variances Efficiency. In V. S. Reddy, V. K. Prasad, D. N. Mallikarjuna Rao, & S. C. Satapathy (Eds.), Intelligent Systems and Sustainable Computing (pp. 59–67). Springer Nature Singapore.
22. Singh, D. G. (2022). Machine Learning Models in Stock Market Prediction. International Journal of Innovative Technology and Exploring Engineering, 11(3), 18–28. https://doi.org/10.35940/ijitee.c9733.0111322
23. Aliyeva, A. (2022). Predicting Stock Prices Using Random Forest and Logistic Regression Algorithms. In R. A. Aliev, J. Kacprzyk, W. Pedrycz, M. Jamshidi, M. Babanli, & F. M. Sadikoglu (Eds.), 11th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions and Artificial Intelligence - ICSCCW-2021 (pp. 95–101). Cham: Springer International Publishing.
24. Demirer, R., Gkillas, K., Gupta, R., & Pierdzioch, C. (2021). Risk aversion and the predictability of crude oil market volatility: A forecasting experiment with random forests. Journal of the Operational Research Society, 1-13.
25. Bouri, E., Gkillas, K., Gupta, R., & Pierdzioch, C. (2021). Forecasting Realized Volatility of Bitcoin: The Role of the Trade War. Computational Economics, 57(1), 29–53. https://doi.org/10.1007/s10614-020-10022-4
26. Gkillas, K., Tantoula, M., & Tzagarakis, M. (2021). Transaction activity and bitcoin realized volatility. Operations Research Letters, 49(5), 715–719. https://doi.org/https://doi.org/10.1016/j.orl.2021.06.016
27. Görgen, K., Meirer, J., & Schienle, M. (2022). Predicting Value at Risk for Cryptocurrencies Using Generalized Random Forests.
28. Suykens, J. A. K., & Vandewalle, J. (1999). Neural Processing Letters, 9(3), 293–300. doi:10.1023/a:1018628609742
29. Dritsas, E., Livieris, I. E., Giotopoulos, K., & Theodorakopoulos, L. (2018, November). An apache spark implementation for graph-based hashtag sentiment classification on twitter. In Proceedings of the 22nd Pan-Hellenic Conference on Informatics (pp. 255-260).
30. Theodorakopoulos, L. (2022). Big data analysis in humanities and economics with machine learning techniques and use of cloud computing technologies. The University of Patras. Department of Business Administration.
31. Antonopoulou, H., Mamalougou, V., & Theodorakopoulos, L. (2022). The Role of Economic Policy Uncertainty in Predicting Stock Return Volatility in the Banking Industry: A Big Data Analysis. Emerging Science Journal, 6(3), 569–577. doi:10.28991/esj-2022-06-03-011
32. Lazzeri, Francesca. 2019b. "Data Science Mindset: Six Principles to Build Healthy Data-Driven Organizations." InfoQ blog. https://www.infoq.com/articles/data-science-organization-framework.