Financial Data Analysis: Forecasting Stock Prices Through Machine Learning

Amit Gaddi gaddiat@mail.uc.edu
School of Information Technology, University of Cincinnati, Ohio, USA

Abstract

The goal of this research is to forecast stock values by utilizing machine learning methods on historical stock market data. Accurately predicting stock fluctuations is difficult because of the complexity and volatility of financial markets. For this goal, several machine learning models have been investigated in earlier research, with variable degrees of success. In this work, we first apply simple, easily understood models such as linear regression as a first step before exploring more complex methods.

The study makes use of a dataset that includes historical stock prices for the top 10 firms. The dataset has undergone extensive preparation procedures, such as feature engineering, normalization, outlier removal, and treatment of missing data. A range of visual aids are utilized to enhance comprehension of the trends and patterns present in the data. Three machine learning models are compared: decision trees, random, and linear regression is used to provide a baseline.

The findings demonstrate that, with R-squared value above 99%, did remarkably well in forecasting closing stock prices based on opening prices. Every model has its own advantages, but linear regression offers simplicity and interpretability. It is noteworthy that the strongest predictor of closing price was opening price.

The results highlight how well historical market data and simple machine learning models work for stock forecasting. Real-time data feeds, sophisticated feature engineering methods, and the investigation of deep learning architectures to improve prediction skills can all be added to this study in the future.

Keywords: Regression, Machine Learning, Stocks, Prediction, Linear Regression

1. Introduction

A key element of the global financial system is the stock market, where values of stocks represent the worth of companies as well as the state of the market. Because precise stock price forecasting has the potential to yield large gains, academics and investors have long pursued this goal. However, because of the market's volatility and the wide range of contributing factors—from corporate performance to economic events—predicting stock movements continues to be difficult [10],[7].

Different machine learning methods have been investigated in the past for stock price prediction, with varying degrees of success. While some studies support the use of decision tree and linear regression models to capture stock movements, others support the use of deep learning techniques like as LSTM networks to handle intricate correlations seen in market data [2]. However, there is still mystery around the optimal model configurations, feature engineering techniques, and input variable choices.

The purpose of this work is to forecast stock market time series data initially using less complex machine learning model like linear regression. The idea is to get a basic grasp of stock price dynamics by concentrating on historical pricing data before moving on to more sophisticated methods. The goal of the study is to provide comprehensive justifications and direction to enable more people to understand stock forecasting techniques in financial data analysis.

This work is important because it has the potential to create stock forecasting models that are dependable and easy to understand. A focus on clarity and simplicity may encourage people to experiment with stock market forecasting and analysis, which would lead to further studies in this area. Positive results could encourage research into more advanced methods and other data sources, such as sentiment analysis and macroeconomic indicators, to improve forecasts.

The study will make use of a collection of historical stock market data for the top 10 firms sourced from reliable financial databases to accomplish these aims. The data will go through a thorough preparation procedure that includes feature engineering, normalization, addressing missing values, and outlier elimination. To clarify trends, patterns, and connections in the data, extensive visualizations will be made.

The following three machine learning models will be compared: random forests for enhanced prediction through ensemble learning, decision trees to capture nonlinear interactions, and linear regression as a baseline [7], [8]. Preprocessed data will be used for the models' training, and to improve performance, optimization and parameter tweaking strategies will be used.

Metrics like Mean Squared Error (MSE) and R-squared (R^2) will be used to assess the performance of the model to determine the generality and accuracy of predictions. Overfitting will be avoided.

Through a methodical approach and a focus on interpretability, this study aims to improve stock market forecasting methods and assist well-informed financial and investment decision-making.

2. Literature review

In the field of financial data analysis, the application of machine learning techniques for stock price prediction has attracted a lot of interest [1]. This review of the literature looks at research that use algorithms such as decision trees, linear regression, and sophisticated models to study the dynamics of the stock market. It assesses how well these strategies handle the complexity of stock prices that is impacted by different causes. Through a critical analysis of the literature, this study seeks to determine the advantages and disadvantages of forecasting techniques, evaluate the importance of data sources, and add to the conversation about how to balance prediction models' simplicity and accuracy.

Predicting stock prices has always been a component of financial study. Conventional techniques such as technical and fundamental analysis concentrate on past patterns and assessments of the firm [4]. These techniques, however, have trouble understanding the complex linkages that drive changes in stock prices.

Data-driven methods for stock forecasting have been made possible by machine learning, which uses algorithms to find intricate patterns. The Adaptive Markets Hypothesis (AMH) contends that market efficiency fluctuates over time, allowing for forecasting, in contrast to the Efficient Market Hypothesis (EMH) which holds that stock prices represent all available information.

Research has looked into using machine learning to forecast stocks. [6] created a system that forecasts US stock movements with an accuracy of up to 77.6% by utilizing temporal connections. When [2] compared deep learning models such as LSTM and GRU with classical techniques, they discovered that the former performed better on bigger datasets.

While [2] preferred ANN over Random Forest for daily closing prices, [1] enhanced forecasts by transforming continuous data to discrete trends. [5] highlighted the efficacy of hybrid models.

The ability of machine learning to identify stock market trends is promising. Research emphasizes the value of feature engineering [1], [3], the efficacy of hybrid models [5], and the promise of deep learning [5]. Nonetheless, there are difficulties with the model's interpretability and the inclusion of external variables.

Improving model interpretability, utilizing a variety of data sources such as news sentiment, creating models for longer-term forecasting, resolving challenges with real-world application, and investigating sophisticated deep learning models should be the top priorities of future research in stock price prediction. By concentrating on these topics, scholars may aid in the creation of prediction models that are more transparent, reliable, and efficient and that better suit real-world investment requirements and market dynamics. The banking and investment industries will be able to make better decisions thanks to these initiatives, which will ultimately improve the accuracy and usefulness of stock forecasting methods.

Although machine learning provides insights into the dynamics of stock prices, issues with data diversity and model openness still exist. Before moving on to more sophisticated methods, this research investigates easily understood models such as decision trees and linear regression to lay the groundwork for comprehension [9]. The objective is to promote a deeper understanding of stock forecasting and aid in the creation of dependable and useful prediction models.

3. Methodology

Dataset acquisition:

A reliable internet resource, Kaggle (https://www.kaggle.com/datasets/khushipitroda/stock-market-historical-data-of-top-10-companies/data), provided the dataset used for this study. The top ten firms' historical stock market data is included, along with details like the company name, date, opening and closing prices, highest and lowest prices, and trading volume. The dataset's dependability and applicability for stock price prediction and research were ensured by its meticulous curation from public financial information and market databases. Figures 1 & 2 represent the companies and their stock prices. And from these trends as the swings of the Netflix were different form others so this study will be using Netflix company for the rest of the study.

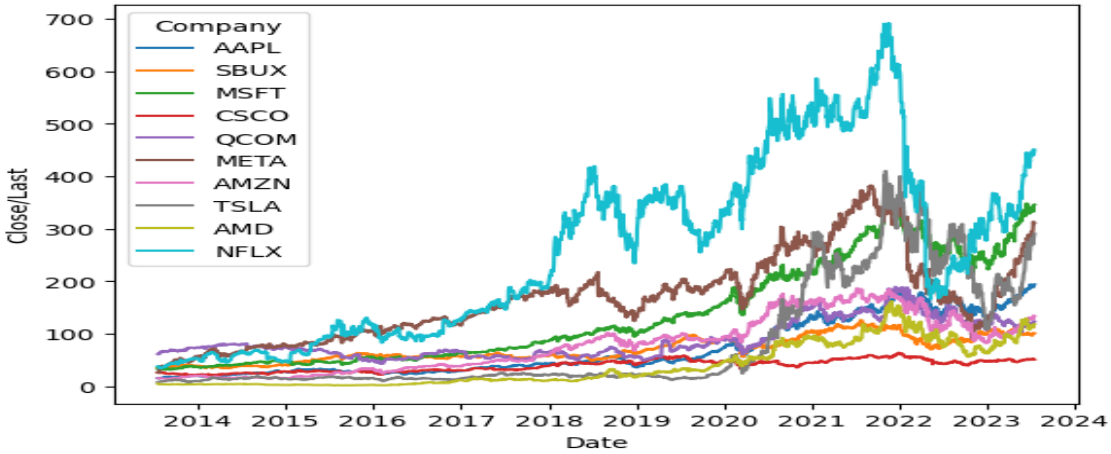

Figure 1. Plot graph of companies share price.

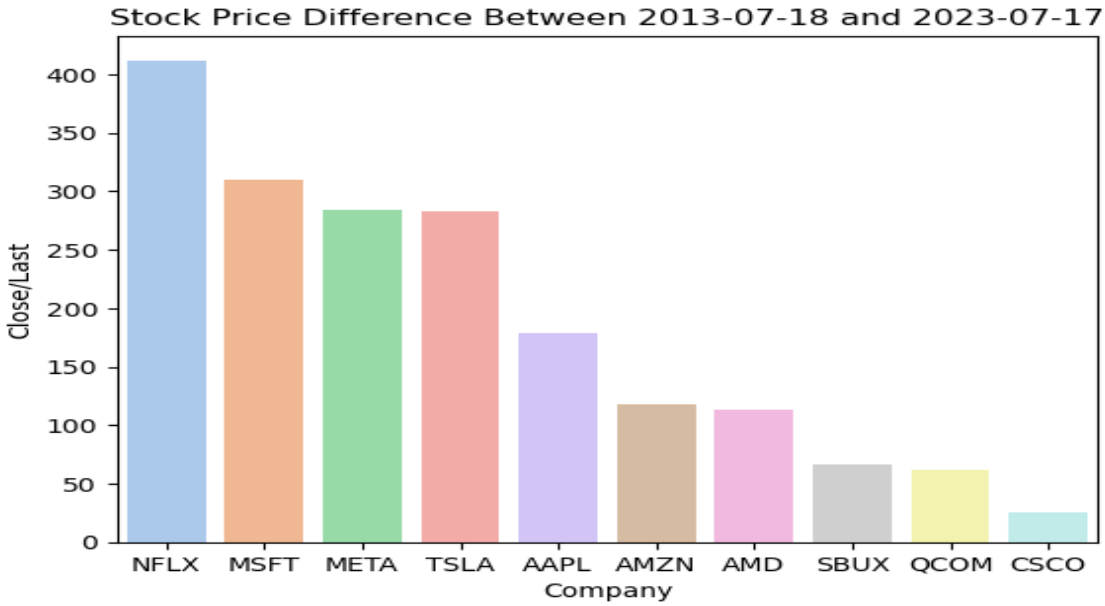

Figure 2. Bar graph of price difference between companies.

Data preprocessing:

A thorough preparation step was performed on the obtained dataset to guarantee its quality and appropriateness for analysis and model construction. Missing numbers, outliers (Figure 3 shows the outlier in the volume for the Netflix company), and inconsistencies were dealt with using a variety of data cleaning strategies. Depending on the context and distribution of the missing data, relevant approaches like interpolation/removal were used to resolve the missing values. To avoid skewing the data, outliers were found and fixed/eliminated using statistical techniques like z-score analysis.

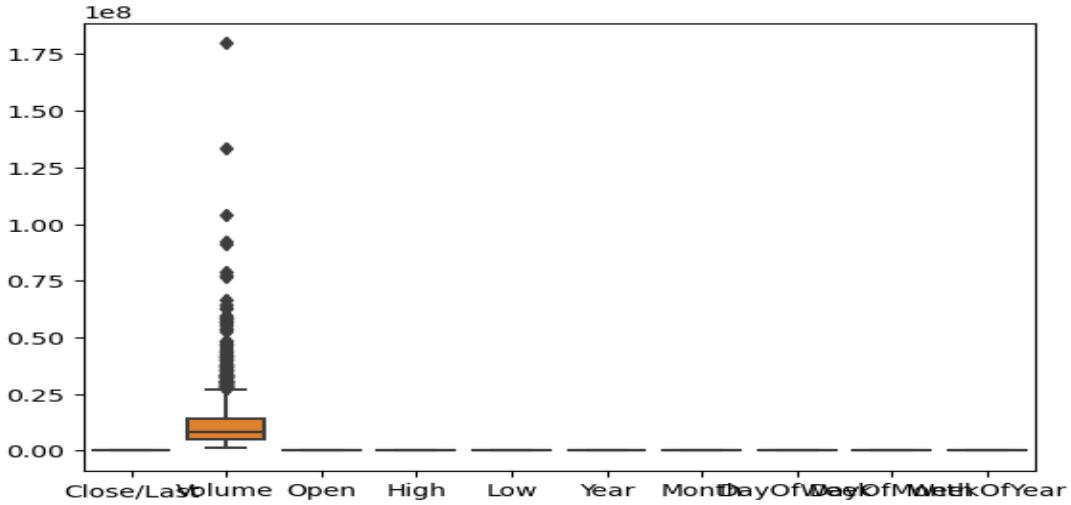

Figure 3. Boxplot outliers' graph (Netflix)

Additionally, data normalization was carried by utilizing methods like Min-Max scaling to scale numerical characteristics to a conventional range, usually between 0 and 1. By ensuring that every feature is on a same size, this step can enhance how well some machine learning algorithms work. To improve the models' predictive capacity, feature engineering was also used to create new features and modify preexisting ones.

Ultimately, training and testing sets were created from the preprocessed dataset, often with 80% of the dataset going toward training and 20% toward testing. Because of this segmentation, the testing set may be used to assess the model's performance on untested data, while the training set is used for model training.

Data visualization:

In this project, data visualization was essential since it made insights from the examined data easier to interpret and share. The value of applying visualization techniques is found in their capacity to improve comprehension, spot trends, provide light on hidden information, and aid in decision-making.

A range of visualization methods were used, such as bar charts, heatmaps, scatter plots, and line graphs. Given their ability to clearly communicate patterns, correlations, and interactions within the dataset, these methodologies were selected considering the study objectives and the characteristics of the data. To create aesthetically pleasing and educational charts, well-known data visualization packages like Matplotlib and Seaborn were used.

Understanding the data was made easier by the visualizations, which showed trends, oscillations, and the relative success of the firms in terms of their stock prices over time. Additionally, correlation heatmaps (Figure 4 showcases the correlation matrix of the Netflix data, from which the study used the 'open' as it's predictor.) were created to show the direction and intensity of correlations between various characteristics. These heatmaps offer important insights for feature selection and help identify underlying patterns in the data.

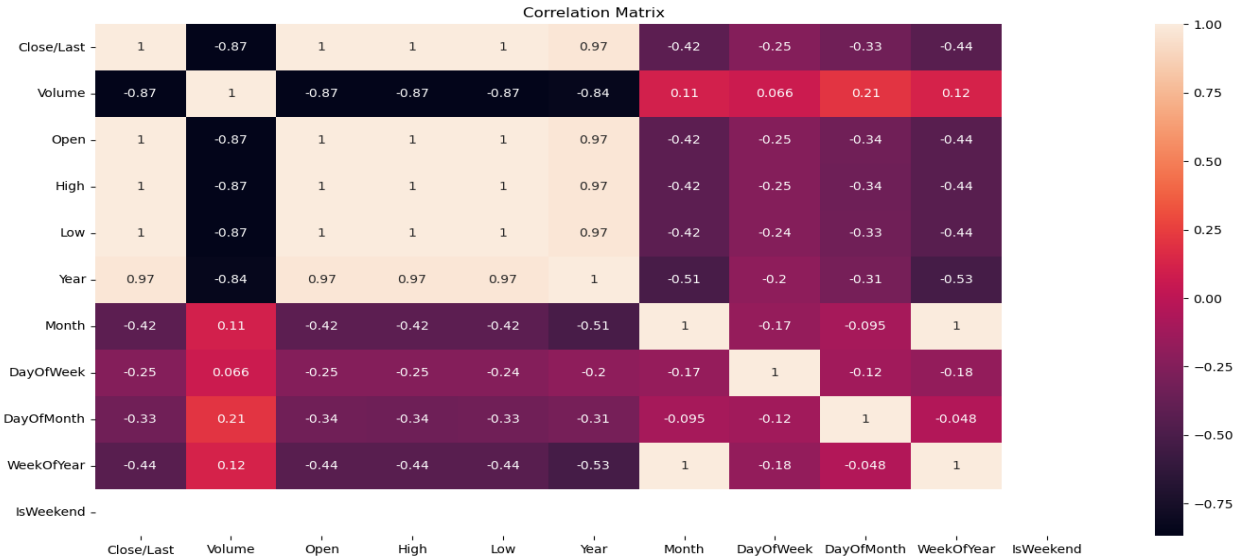

Figure 4. Correlation matrix heat map (Netflix)

Model selection and training:

To bolster the hypothesis of stock price prediction, three machine learning models—linear regression, decision tree regression, and random forest regression—were chosen. These models were selected due to their ease of use, interpretability, and efficiency with time-series data, such as stock prices.

1. Linear Regression: A basic model used to provide a starting point for forecasting based on past stock values. It is predicated on the predictor factors (such opening prices) and the target variable (closing prices) having a linear connection.

2. Decision Tree Regression: A comprehensible model appropriate for expressing feature interactions and non-linear correlations. Decision trees work well with complicated datasets because they recursively divide the data according to feature values [7].

3. Experiential Forest Regression: a strong ensemble learning method that minimizes overfitting and increases prediction accuracy by combining many decision trees. The predictions of several trees trained on various data subsets are combined by the random forest model [8].

The preprocessed dataset was used to train these models, and to improve performance, parameter adjustment and optimization were carried out. The ideal hyperparameters for each model were determined using strategies like grid search and randomized search, guaranteeing the greatest possible fit to the data, the authors for Decision Tree and Forest have used different dataset.

Performance evaluation:

The trained model accuracy and capacity for generalization were tested by the use of suitable measures to evaluate their performance. The following measurements were used:

1. The MSE (mean square error): This metric gives information on the prediction accuracy of the model by calculating the average squared difference between the actual and predicted values. Better model performance is indicated by lower MSE values.

2. R^2, or R-squared: The percentage of the target variable's variance that can be accounted for by the predictor variables is shown by the R-squared score. A model that fits the data better is indicated by an R-squared value that is closer to 1.

These metrics were selected due to their applicability to regression issues and their capacity to measure the performance of the models in terms of variance explained and prediction accuracy.

Throughout the assessment process, each model's performance was compared across the selected metrics, their advantages and disadvantages were examined, and judgments on each model's applicability for stock price prediction were reached. This thorough assessment process the study tries to determine that the best model is chosen for the given dataset and issue.

4. Results and discussions

Summary of Findings:

Using machine learning models to analyze historical stock market data produced some interesting results. The following is a summary of the main findings:

Model Performance: With high R-squared value above 99%, Linear Regression performed very well in forecasting stock closing prices. This suggests that the fundamental relationships and patterns in the data were successfully captured by the models. Figure 5 represents the actual and predicted plot the study.

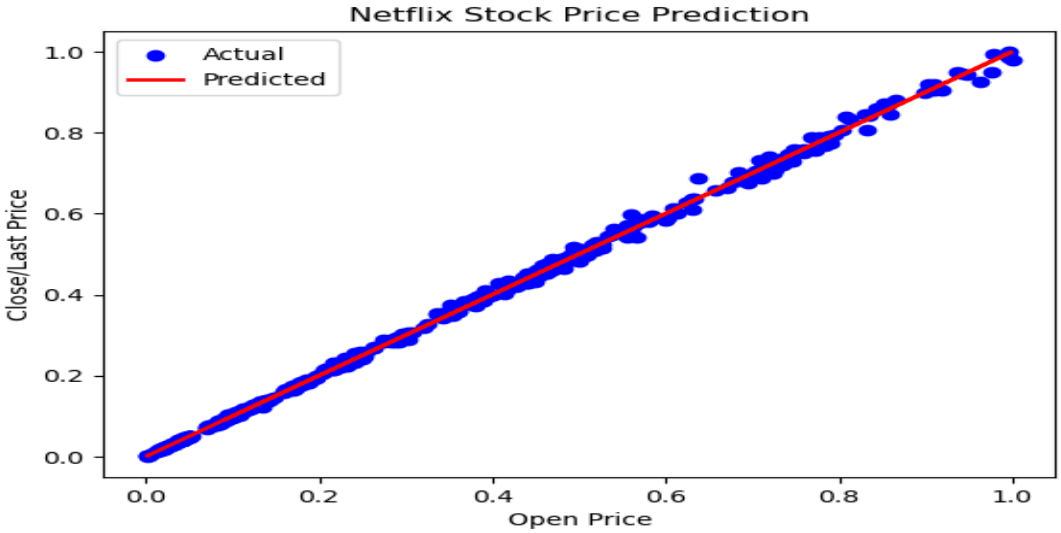

Figure 5. Netflix Actual vs Predicted graph.

Feature Importance: The strongest link seen during data preprocessing was confirmed by the 'Open' price, or opening price, which turned out to be the most important predictor of closing prices. The target variable was less affected by other characteristics, such as trade volume and temporal factors.

Comparative Analysis: The models showed a range of advantages and disadvantages. Decision Tree Regression identified non-linear correlations, Random Forest Regression minimized overfitting, and Linear Regression offered easily interpreted coefficients and simple forecasts.

Explanation of Results:

The significance of utilizing previous opening prices (also known as "Open") as a leading indication for projecting closing prices (also known as "Close/Last") is shown by the models' excellent predictive accuracy. The models' strong performance indicates that they made good use of this feature to predict stock prices. Opening prices are the most important factor in stock price prediction, as evidenced by the little impact of other characteristics such as trading volume and temporal factors on the target variable.

Opening and closing prices have a significant positive association, according to the models' interpretation of the observed trends and patterns in the data. Investors and analysts may use this information to help them make well-informed judgments about investing methods and stock trading.

Comparative Analysis:
Subtle variations between the models were discovered through comparison:

Decision Tree vs. Linear Regression: The decision tree model performs better in terms of trading strategy effectiveness (ACARR) when compared to the linear regression model. Decision trees are better suited for tasks involving trade point categorization or grouping because they are skilled at capturing complicated relationships and non-linear patterns within the data. On the other hand, whereas linear regression offers a simple way to understand relationships, it might not be sufficient when data shows non-linear behavior.

Linear Regression vs. Random Forest: Because random forests can handle complicated data structures and represent non-linear connections, they perform much better in forecasting stock price directions than linear regression. While random forests use ensemble learning to attain improved accuracy, particularly for classification tasks like stock price direction prediction, linear regression provides transparency and simplicity of understanding. Technical indicators are added to the model, which improves its performance even further and makes it a strong option for stock market predictions.

The comparison emphasizes how crucial it is to choose a model depending on the particular goals of the project and the properties of the dataset. For this investigation, the transparency of Linear Regression was sufficient, while Decision Tree and Random Forest models provided more in-depth understanding of complex data linkages.

Comparison with Expectations or Prior Knowledge:
The results are in line with predictions and support the widely held belief that starting prices have a big impact on closing prices in stock markets. This premise is supported by the models' accuracy in predicting closing prices, which offers empirical evidence in favor of well-established financial theories.

Addressing Research Questions
The findings specifically address the goal of the study, which was to forecast stock prices using historical data. The project's objectives are greatly advanced by the models' remarkable performance and feature significance analysis. The findings give stakeholders in the financial and investment industries useful information by clarifying the opening price's predictive capacity.

Limitations and Uncertainties:
Even if the models performed well, there are still certain issues that need to be addressed:

Data Quality: Because the analysis is based on historical data, it is prone to biases and problems with data quality.
Feature Selection: Removing some characteristics might reduce the prediction power of the models, thus more research is necessary.
Model Complexity: Although strong models, Random Forest and Decision Tree models need meticulous parameter adjustment to maximize efficiency.

These restrictions point to areas that need more investigation and prediction model improvement.

Visual Aids:
The comprehension of results is improved by the use of visual aids such feature significance plots and line charts that compare actual and forecasted closing prices. In order to help with a better comprehension of the data patterns, these graphics offer understandable representations of feature influences and model predictions.

Implications and Significance:
The results have important ramifications for financial academics, analysts, and investors. Reliability in stock price prediction based on opening prices has useful implications for trading techniques, risk management, and portfolio

management. Additionally, the work advances our understanding of predictive modeling in stock market analysis and adds to the larger conversation on machine learning applications in financial forecasting.

5. Conclusions and future work

Finally, based on previous opening prices, this study effectively illustrated the predictive power of machine learning models, particularly Linear Regression, Decision Tree Regression, and Random Forest Regression. The results of the investigation showed that opening prices, or "Open," are an important predictor of closing prices, or "Close/Last." The models demonstrated an impressive level of predictive accuracy, with R-squared values reaching 99%.

The following are the main ideas from this project summarized:

Importance of Feature: The most significant factor was the 'Open' price, which demonstrated a substantial relationship between it and closing prices.

- Model Effectiveness: Different models demonstrated different strengths: Random Forest Regression balanced model complexity and generalization, Decision Tree Regression captured non-linear correlations, and Linear Regression provided transparency.

The results highlight the importance of using previous market data to guide trading strategies and investment decisions. They also highlight the importance of opening prices in predicting closing prices.

Limitations and Possible Future Work:

Notwithstanding the models' effectiveness, there are a few drawbacks and directions for further study that should be taken into account:

- Data Quality: More research on biases, missing features, and problems with data quality might improve the robustness of the model.

- Feature Development: Adding more information and experimenting with different transformations might reveal undiscovered connections and enhance prediction accuracy.

- Model Enhancement: Refinement of model parameters and tuning methods might be pursued to maximize generalizability and prediction accuracy.

- Market Structure: Predictive capacities may be improved by using real-time market data and external elements (such as news mood and economic indicators).

Furthermore, investigating different ensemble techniques and machine learning methodologies may enhance the field of predictive modeling. Applying deep learning architectures, including Long Short-Term Memory (LSTM) networks, may improve predicting accuracy and provide deeper insights into temporal connections.

To summarize, this study establishes a solid basis for utilizing machine learning methods in financial forecasting. Going forward, research efforts will focus on resolving existing constraints and enhancing the predictive modeling capacities in stock market analysis. Through tackling these obstacles, scholars and professionals might uncover novel perspectives and uses in the ever-changing field of financial markets.

References

[1] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.

[2] A. Site, D. Birant and Z. Işık, "Stock Market Forecasting Using Machine Learning Models," 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 2019, pp. 1-6, doi: 10.1109/ASYU48272.2019.8946372.

[3] Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia computer science, 167, 599-606.

[4] Sonkavde G, Dharrao DS, Bongale AM, Deokate ST, Doreswamy D, Bhat SK. Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. International Journal of Financial Studies. 2023; 11(3):94. https://doi.org/10.3390/ijfs11030094

[5] S. Singh, T. K. Madan, J. Kumar and A. K. Singh, "Stock Market Forecasting using Machine Learning: Today and Tomorrow," *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Kannur, India, 2019, pp. 738-745, doi: 10.1109/ICICICT46008.2019.8993160.

[6] Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department* of Electrical Engineering, Stanford University, Stanford, CA, 1-5.

[7] Wu, M. C., Lin, S. Y., & Lin, C. H. (2006). An effective application of decision tree to stock trading. Expert Systems with applications, 31(2), 270-274.

[8] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

[9] Siew, H. L., & Nordin, M. J. (2012, September). Regression techniques for the prediction of stock price trend. In 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE) (pp. 1-5). IEEE.

[10] Leung, C. K. S., MacKinnon, R. K., & Wang, Y. (2014, July). A machine learning approach for stock price prediction. In Proceedings of the 18th international database engineering & applications symposium (pp. 274-277)

Appendix:

Disclaimer:

This study's forecasts and assessments are predicated on historical stock market data using machine learning models, such as random forests, decision trees, and linear regression. Although these techniques have proven successful in specific situations, future investment performance or results cannot be guaranteed. The stock market is naturally volatile and susceptible to a number of factors outside the purview of statistical models, such as changes in the economy, developments in geopolitics, and mood inside the market. Consequently, it is not appropriate to use the study's conclusions as suggestions or financial advice while making investment decisions. Before considering an investment, investors are advised to perform extensive research, evaluate their own risk tolerance, and speak with licensed financial advisors. The study's authors disclaim all liability and responsibility for any investment decisions made using the information it contains.