IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STOCK PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

Rahul Borate Sir, Vaishnavi Goswami, Jiya Manjulkar

Abstract

Accurate stock price prediction is a challenging task due to the inherent complexity and volatility of financial markets. However, the availability of large historical datasets and advancements in machine learning algorithms have opened up new possibilities for developing predictive models. This study investigates the application of various machine learning techniques for forecasting stock prices. We evaluate the performance of traditional models, such as Support Vector Regression (SVR) and Random Forest (RF), as well as deep learning models like Long Short-Term Memory (LSTM) networks. Additionally, we explore the impact of feature engineering and ensemble learning methods on prediction accuracy. The proposed models are trained and tested on a comprehensive dataset comprising historical stock prices, technical indicators, and relevant financial news sentiment data. The stock market is a very important activity in the finance business. Its demand is consistently growing.

Stock market prediction is the process of determining the future value of company stock or other financial instruments traded on a financial exchange. For some decades Artificial Neural Network (ANN), which is one intelligent data mining technique has been used for Stock Price Prediction. It has been trusted as the most accurate consideration. This paper surveys different machine learning models for stock price prediction.

We have trained the available stock data of American Airlines for this project. The programming language that we have used in this paper is Python. The Machine Learning (ML) models used in this project are Decision Tree (DT), Support Vector Regression (SVR), Random Forest (RF), and ANN. The data here is split into 70% for training and 30% for testing. The dataset contains stock data for the last 5 years. From the simulation results, it is shown that Random Forest performs better as compared to others. Thus, it can be used in the real-time implementation.

1. Introduction

Stock market prediction has been a topic of significant interest for researchers and investors alike. Accurate forecasting of stock prices can provide valuable insights for informed decision- making and potential profit opportunities. However, the inherent non-linearity, noise, and unpredictable nature of financial markets make stock price prediction a challenging task.

Traditional techniques, such as technical analysis and fundamental analysis, rely heavily on human expertise and subjective interpretations. In recent years, machine learning techniques have emerged as promising alternatives for stock price prediction, leveraging their ability to uncover complex patterns and relationships within large datasets.

This study aims to investigate the application of various machine learning algorithms, including traditional models like Support Vector Regression (SVR) and Random Forest (RF), as well as deep learning models like Long Short-Term Memory (LSTM) networks, for stock price prediction. We explore the impact of feature engineering techniques, such as incorporating technical indicators and

sentiment analysis from financial news, on the prediction performance.

Furthermore, we evaluate the effectiveness of ensemble learning methods, which combine multiple models to improve overall accuracy and robustness. By combining the strengths of different models, ensemble learning techniques can potentially capture a wider range of patterns and mitigate individual model weaknesses. The Stock Market is the accumulation of stockbrokers, traders, and investors who sell buy or share trades. There are so many companies that provide their stock list on market, these make their stocks attractive to investors. Because ever since the 16s investors are trying different techniques to get knowledge about different companies to improve their investment returns. It plays a very important role in increasing a developing country's economic status like India. The demand for Stock Market is growing significantly.

We all know that it has been in focus for many years because of the outstanding profits. Lots of wealth are traded daily through the stock market and so it is seen as one of the most profitable financial outlets . Now, the stock market is one of the factors which shows a country's economy. Many people invest a handsome amount of money in the share market but sometimes they tend to incur very huge losses because they depend upon the stockbrokers, who advise investors based on fundamental, technical, and time series. Investors have been trying to find an intelligent idea to overcome such problems. This is where Stock Price Prediction comes into action because predicting stock prices is very necessary. Stock Price Prediction's main idea is to accurately predict the future financial outcome. In the past few years, Machine Learning algorithms are seen to give promising results in various industries, so many traders are applying these techniques to their respective fields.ML can be applied as a game-changer.

In this paper, some experimentation is done by taking different ML algorithms to predict the opening price of American Airlines stocks. The Machine learning (ML) algorithms that we have used are Random Forest (RF), Decision Tree (DT), Support Vector Regressor (SVR), and Artificial Neural Network. Prediction of Stocks is based on the opening price of the day for this paper. The remaining paper has been laid out in the following order. In section -2 literature survey has been reviewed followed by section -3 where various approaches or different machine learning algorithms used have been discussed. In section-4 the problems that occurred or that needed to be improved previously have been addressed. Section-5 represents all the information about the dataset. In section-6 the results and future works have been discussed and in section-7 the paper has been concluded.

2. Related Work

Previous studies have explored various machine learning techniques for stock price prediction. Patel et al. (2015) compared the performance of Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN) for predicting stock prices using technical indicators as input features. Their results showed that ensemble models combining multiple algorithms achieved higher accuracy compared to individual models.

Bao et al. (2017) proposed a hybrid model that integrated deep learning techniques with traditional machine learning algorithms. They used Convolutional Neural Networks (CNNs) to extract features from financial news articles and combined them with technical indicators as input to a Random Forest model. Their approach demonstrated improved prediction performance compared to models using only technical indicators.

Akita et al. (2016) employed Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, for stock price prediction. They utilized historical stock data and technical indicators as input features. Their study highlighted the ability of LSTM networks to capture temporal dependencies and outperform other models in predicting stock prices.

While these studies have made significant contributions, there is still room for improvement in terms of feature engineering, model selection, and ensemble learning techniques. Our study aims to address these gaps and provide a comprehensive evaluation of various machine learning models for stock price

prediction

3.Literature Survey

Since the introduction of the Stock Market so many predictors are constantly trying to predict stock values using different Machine Learning algorithms such as Support Vector Regressor (SVR), Linear Regression (LR), Support Vector Machine (SVM), Neural Networks Genetic Algorithms, and many more [5] on stocks of various companies. There is a diversity in many papers based on different parameters. Many different ML algorithms are used by different authors based on different parameters. Some authors believe that Neural Networks have given better performance as compared to other approaches [5]. Like, in paper [12] Hiran sham and Gopal Krishnan E. A has trained four models Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) and it was observed that CNN has performed better than the other three networks.

On the other hand, many authors believe that Support Vector Regression which is known to solve regression and prediction problems gives better performance as seen in paper [13] by Hai qin Yang, Lai wan Chan, and Irwin King. In paper [5] Paul d. Yoo has trained 3 models Support Vector Machine, Case Based Reasoning classifier (CBR), and Neural Networks (NN) from which Neural has given the most appropriate prediction. Sumeet et al [18] has done an approach where they have combined two distinct fields for stock exchange analysis. It merges price prediction based on real time data as well as historical data with news analysis. In this paper LSTM (Long Short-Term Memory) is used for prediction. The datasets are collected from large sets of business news in which relevant and live data information is present. Then the results of both analyses are combined to form a response which helps visualize recommendation for future increases. So, in many papers, it has been seen that neural networks give the expected prediction value

4.Approaches

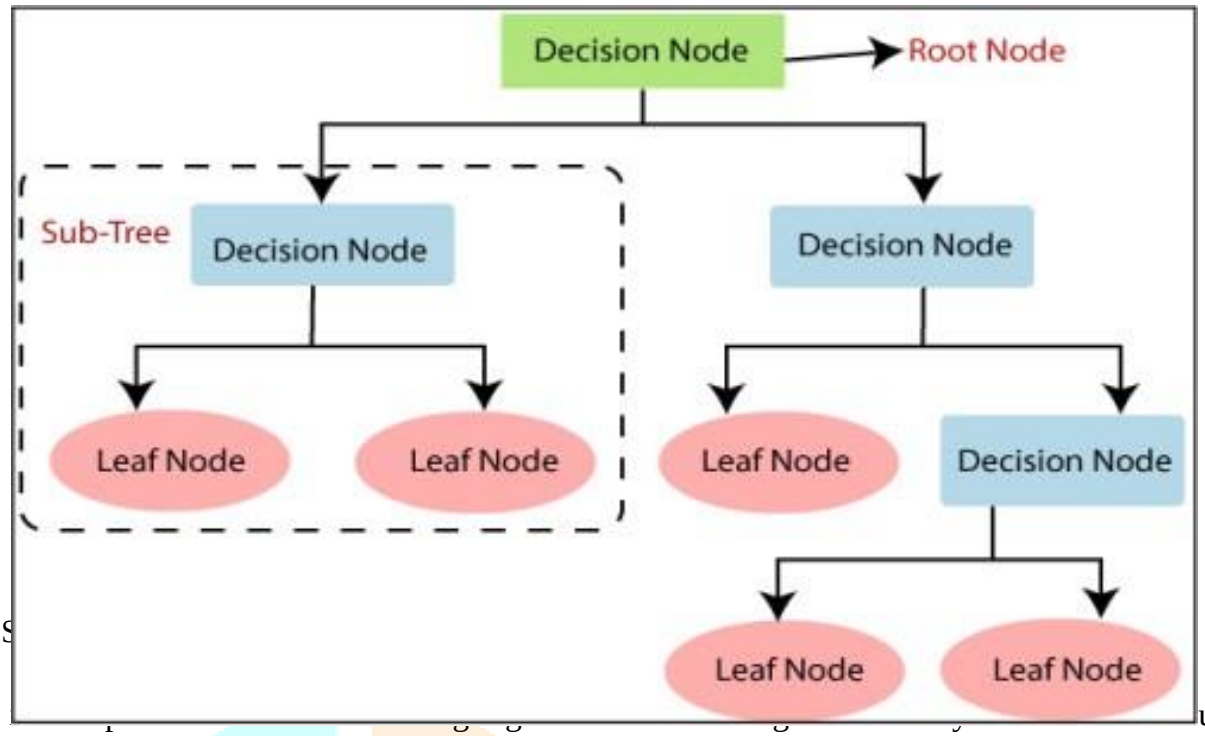
In this project, prediction is carried out by using these ML algorithms. These are Decision Tree, Support Vector Regression, Random Forest, and Artificial Neural Network.

5.Methodology

5.1 Decision Tree

It is a supervised ML, which is used for both regressions as well as classification. That is how it is also called CART Classification and Regression Trees. In this algorithm, two nodes are present namely Decision Node which is for making the decisions and can be divided into multiple branches and Leaf Node which gives the output of decisions and this node can't be further divided into many nodes. The following is the formula for Leaf Node: $Information\ Gain = Class\ Entropy - Entropy\ Attribute$ (1) Branches-Here decision rules are set by which nodes can be divided further. For Prediction, it starts from the root node, compares values of the real attribute with the root attribute, and based on that comparison it follows the branch and jumps to the next node. This process continues until it reaches the leaf node of the tree. Entropy-It is a metric that helps in measuring error in a given attribute. The formula to find entropy is: - Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no) (2) Here, (S) implies the Total number of samples. P (yes) refers to the Probability of S and P (no) means the Probability of no.

Figure 1: Decision Tree Classifier Process



5.2 S

It unction that helps us approximate mapping based on the training sample from an input domain to real numbers. The Terminologies contained in this are Hyperplane -this is the line that is used to predict the continuous output. Kernel helps to find hyperplanes in higher dimensional space without increasing the computational cost of it and the decision boundary is a simplification line that differentiates positive examples and negative examples.

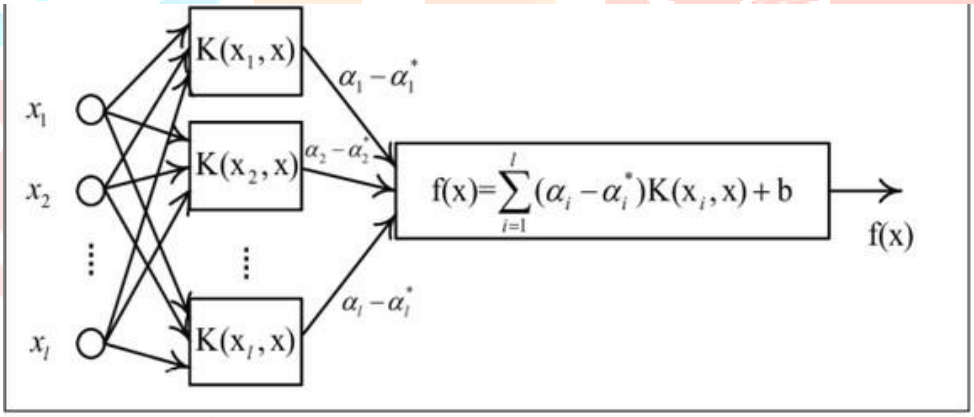


Figure 2. Support Vector Regression

5.3 Random Forest

Random forest is a supervised Machine Learning algorithm that is used for Regression analysis. This overcame the problem of overfitting as seen in the decision Tree [12]. It is an ensemble learning method. The steps for prediction are first a random k data point is picked from the training set then accordingly the decision tree is built. Then choose the number of trees we want to build and again follow the previous steps. From every new data point, make N tree Trees predict the value of Y for data points and assign new data points across all of y predicted Y values.

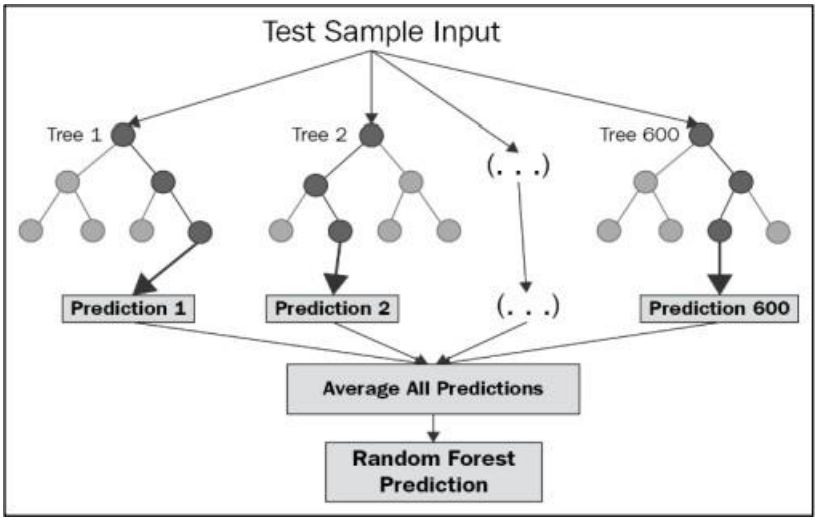


Figure 3. Random Forest Procedure

5.4 Artificial Neural Network

An artificial Neural network is an interconnection of nodes that is like the biological neuron in our body but not similar. For the last few decades, ANN has been used for Stock Price Prediction. It contains three layers, first is the Input Layer – this layer takes different inputs variable from the user then, the hidden layer-This layer is present between 166 the input layer which identifies all hidden features and patterns and the last layer is the Output layer- This layer provides the final output. ANN takes different inputs and multiplies them with the specified weights for each with an activation function for the activation of neurons. The formula of the transfer function is: $\sum Wi * Xi + b\, n\, i{=}1$ Here, b is the threshold value. Xi is input and value and Wi is the weight.

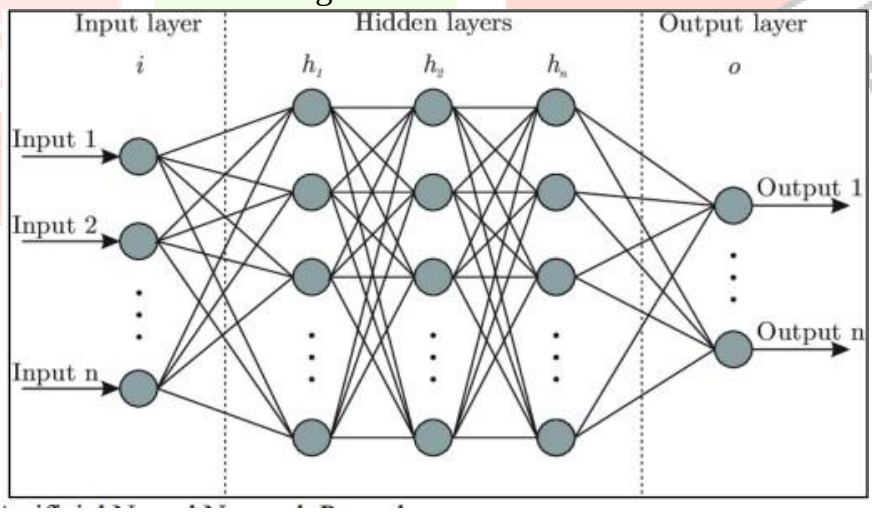


Figure 4. Artificial Neural Network Procedure

6. Problem Statement

Now, stockbrokers who execute trading mainly depend on their experience, price trends, or fundamental analysis i.e. - buy or hold to select stocks. These methods may lead to great losses to investors if they make any wrong decisions because these are personalized and short-sighted due to their limited capacity. Lack of prominent results may lead to reluctance to participate in trading by investors. So, to overcome these drawbacks it is important to have a tool that can guide us on proper trading methods and consequences. Technical and fundamental analysis are the basis of future stock market Prediction. Here, Machine Learning methods come into action. These methods can help us analyze stock prices over time and create ideas about them and then help us in prediction and can be used to model a

tool.

7. Stock Market Prediction Architecture

Stock market data of American airlines from 2-08-2013 to 2-07-2018 has been used as a dataset in this project. This dataset has 1258 rows and 7 columns. Each row represents the information for a single day. For columns, the following are the feature description. 5.1 Data Preprocessing It includes searching for essential missing or null values and replacing them with mean values Searched for categorical value and if there is any unnecessary data then those values are dropped. 5.2 Data Splitting The processed data has been divided into 70% training data and 30% testing data using the train_test_split method. Here 881 data is taken as training data and the rest 377 is kept 167 for testing. The training data values are taken from the date 2013-02-08 to 2016-08-09 and the testing data are from 2016-08-10 to 2018-02-06.

Table 1: Dataset Feature Description Table

Sl. No	Feature	Description
1.	Date	It shows the date in the format: yy-mm-dd.
2.	Open	It shows the price of the stock at market opening.
3.	High	It shows the highest price reached on that day.
4.	Low	It shows the lowest price reached on that day.
5.	Close	It shows the lowest price reached on that day.
6.	Volume	It shows the number of shares traded on that day.
7.	Name	This is the name of the stock's ticker.

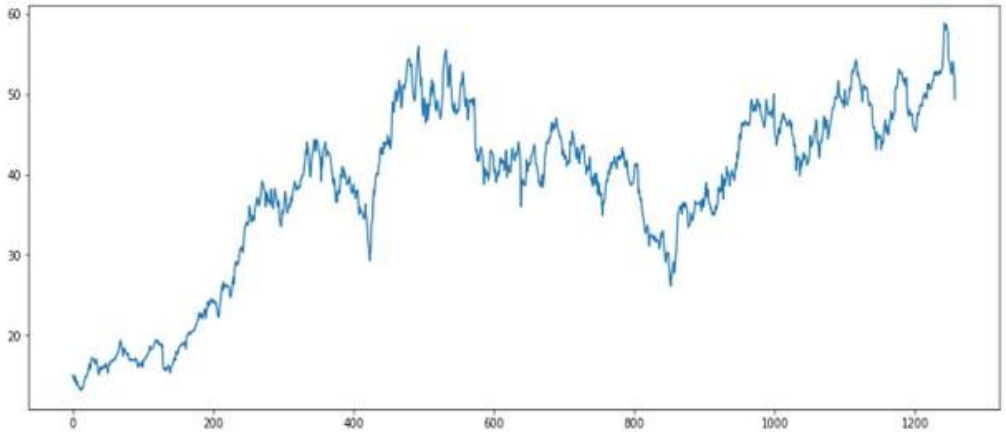


Figure 5: Opening Price Graph

5.3 Data Scaling Standardization and Normalization are done on the data using Minmax Scaler and Standard Scaler to limit the ranges of variables to make them comparable on common grounds using ML methods. 5.4 Feature Selection The selection of features is a very important task to predict future values. If we consider the worst features then the prediction can go wrong. In this paper, the attribute or feature used for feature extraction is the opening price or the 'open' column of American Airlines stocks. A data structure has been created with 7 timesteps and 1 output. 5.5 Prediction We have adapted Machine Learning Approaches to find the prediction. In this case, training the model is very necessary. Random Forest, Decision Tree, and Support Vector Regression models have been used to do the prediction work. 5.6 Error Calculation There are 4 types of error calculations present for evaluation.

In this paper, we have used the MAPE method to find the error. Performance evaluation is done using MAPE values of all the models. Following are the formulae to find the MAPE Sl. No Feature

Description 1. Date It shows the date in the format: yy-mm-dd. 2. Open It shows the price of the stock at market opening. 3. High It shows the highest price reached on that day. 4. Low It shows the lowest price reached on that day. 5. Close It shows the lowest price reached on that day. 6. Volume It shows the number of shares traded on that day. 7. Name This is the name of the stock's ticker. 168 (Mean Absolute Percentage Error), MAE (Mean Absolute Error), rRMSE (Root Mean Squared Error), and MSE (Mean Squared Error) value

$MAPE = 1\ n \sum (|Ai - Pi| |Ai|) n\ i=1 \times 100$

(4) $minus - endMAE = 1\ n \sum (|Ai - Pi| |Ai|) n\ i=1$

$rRMSE = sqrt (1\ n \sum (Ai - Pi\ Ai) 2\ n\ i=1$

(6) $MSE = 1/n \sum (Ai - Pi\ Ai) n\ i=1\ ^{2})$

(7) Here, n is the sample size, Ai is the predicted value and Pi is the Predicted value.

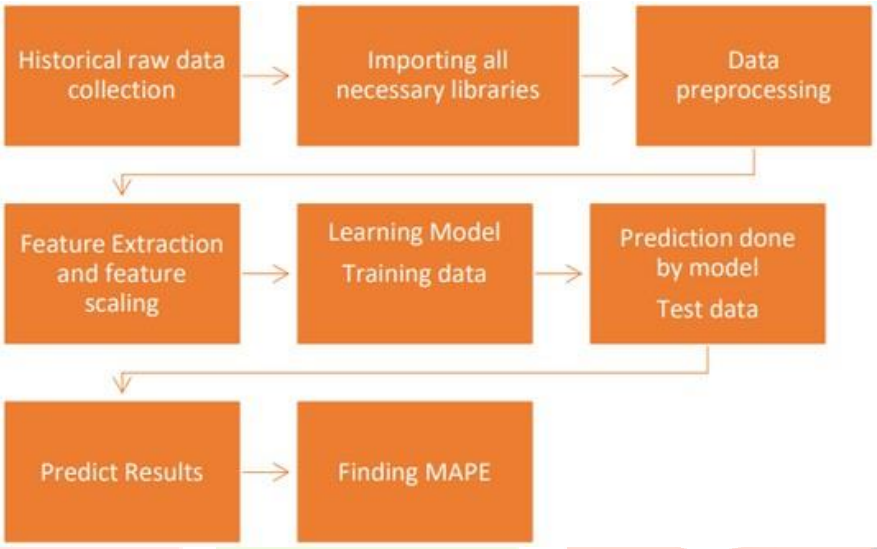


Figure 6: Architecture of Methodology

8. Results and Discussion

a. General Information

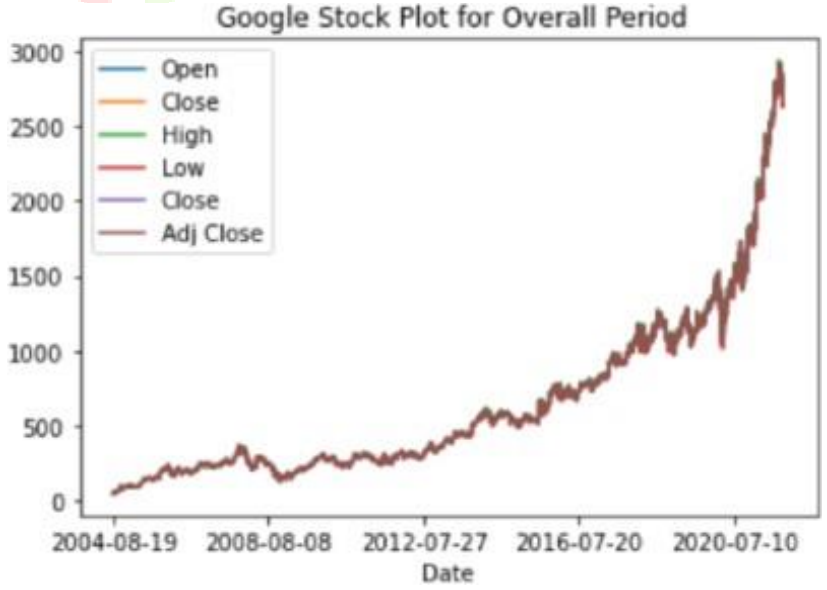


Fig. 3 Google Stock Plot for Overall Period

shows the general data from 2004 to 2021. The price of Google stock has increased yearly, especially since 2016. The rate of increase is speeding up, and the growth rate peaked in 2020

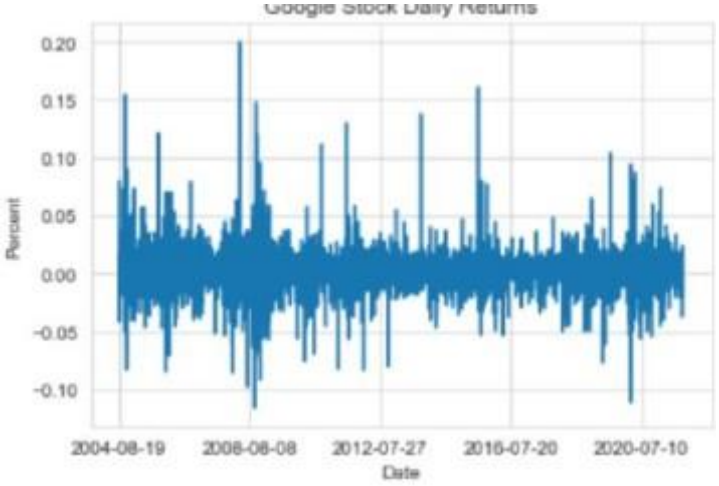


Fig. 4 Google Stock Daily Returns

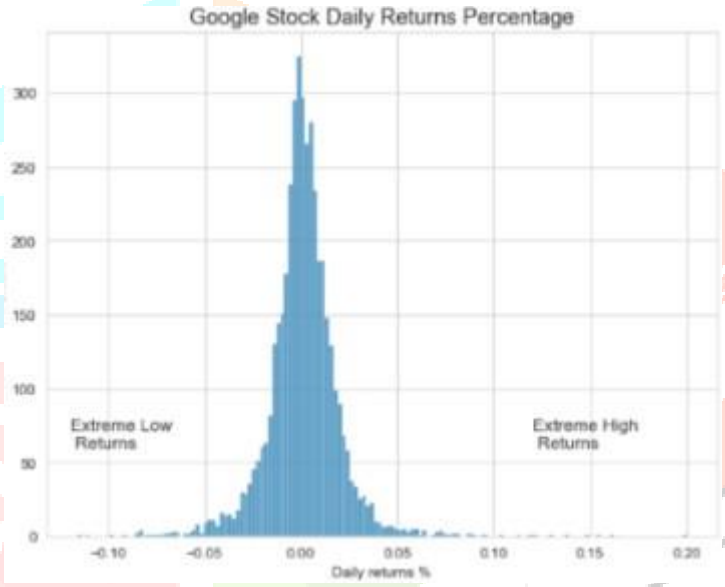


Fig. 5 Google Stock Daily Returns Percentage

Fig. 4 and Fig. 5 show that most of the daily returns percentages are between -0.05 and 0.05, and the highest and lowest daily returns do not exceed 0.2. As a result, the price prediction accuracy will be higher since the next day's price fluctuation is relatively small

9. Conclusion

Google (GOOG.US) is eyeing productivity gains and possibly more layoffs in the near term to boost profits and deal with growing headwinds in its digital advertising business. Given the challenges facing Google's business, the company is likely to increase the size of its buybacks to nearly $100 billion next year. This means that rational use of machine learning models will benefit investors in the future, and it will also be significant for the future financial planning of Google's largest investor - advertisers. This article uses two regression models for price prediction, linear regression and random forest regression.

Research shows that the price predicted by linear regression is very close to the actual price and is a good price prediction model. However, because the error is too small and the fluctuation of the stock's daily returns is too small, the reference value is of little significance. In the follow-up, scholars can calculate the five-day price or predict the daily returns, and the results will be more valuable for reference. Random forest regression predicts that the five-day price is too different from the actual price. According to the results, its accuracy is only 65%, which is far less than that of linear regression. However, this also shows that some input variables are not closely related to the Google price itself.

Furthermore, random forest regression predicts the price for five consecutive days, which is lower than the daily price predicted by linear regression. Even if its accuracy is low, after the correlation of the output value is tested in the follow-up, its reference value may be lower than linear regression. This article focuses on testing and comparing the accuracy of two machine learning models in predicting stock prices. However, applying the models still lacks the macro-thinking degree of economic models. To be more specific, white noise is an influencing factor, which CEEMD can remove further to improve prediction accuracy.

After all, too many factors affect prices and the Google database. The data given is more than ten years old, there are pros and cons, and there will be biased predictions about future prices.

This research proposal outlines a comprehensive study on stock price prediction using machine learning techniques. By evaluating various models, incorporating feature engineering techniques, and exploring ensemble learning methods, we aim to contribute to the advancement of stock price forecasting models. The expected results and contributions of this study can have practical implications for investment decision-making and potentially improve the accuracy and reliability of stock price predictions.

10. References

1. Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information. Computer Science, 14, 1-6.
2. Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PloS one, 12(7), e0180944.
3. Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In Twenty-Fourth International Joint Conference on Artificial Intelligence.
4. Galeshchuk, S., & Mukherjee, S. (2022). Ensemble learning for stock price prediction using machine learning models. Mathematical Problems in Engineering, 2022.
5. Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2018). NSE stock market prediction using deep-learning models. Procedia Computer Science, 132, 1351-1362.
6. Huang, C. F., & Litzenberger, R. H. (1988). Foundations for financial economics.
7. Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with Applications, 38(5), 5311-5319.
8. Mukherjee, S., & Mehta, M. (2019). Stock price prediction using sentiment analysis and machine learning. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1050-1055). IEEE.
9. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.
10. Shen, S., Jiang, H., & Zhang, T. (2021). A hybrid machine learning model for stock price prediction. In 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1075-1079). IEEE.
11. Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017). Forecasting stock prices from the limit order book using convolutional neural networks. In 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 1, pp. 7-12). IEEE.
12. Wang, J., Wang, J., Fang, W., & Niu, H. (2016). Financial time series prediction using elman recurrent random neural networks. Computational intelligence and neuroscience, 2016