# The Prediction of Google Stock Closing Price Based on Linear Regression Model and Random Forest Model

Zixuan Luo
*Guangdong University of Technology, Guangzhou, China*

Keywords: Linear Regression, Random Forest, Stock Price Prediction.

Abstract: Price prediction in the stock market has always been a matter of great concern. Due to the unstable and nonlinear nature of the stock market, predicting stock prices is a very challenging task. To improve the efficiency of stock price prediction, many machine learning algorithms and deep learning models have been developed. These machine learning models have better performance compared to traditional prediction methods. In this study, the stock prices of Google Inc. for the last five years downloaded from Kaggle website are used as experimental data. Linear regression model and random forest model are used to predict the closing price of Google Inc. and are evaluated and compared using three different metrics. The results show that these two machine learning algorithm models are effective in predicting the closing price of a stock and that the linear regression model performs better than the random forest model in the given cases.

## 1 INTRODUCTION

Since the birth of the stock market, financial scientists and sociologists all over the world have customarily taken the degree of development of the stock market as one of the evaluation indexes when evaluating the prosperity and level of development of a certain country or a certain region, so it is clear that the study of the stock market is of practical significance. Price prediction in the stock market has always been a great concern to people and a very challenging task in itself. Because the stock market is characterized by instability and non-linearity, and its formation mechanism is quite complex. Stock price volatility is the result of a combination of factors. Frequent stock price fluctuations amplify speculative activities in the stock market, as speculators tend to take advantage of short-term price fluctuations to make profits. Such speculative activities increase the risk of the stock market, as price fluctuations may lead to losses for speculators. In addition, speculative activities may lead to increased market instability and volatility, creating uncertainty and risk for other investors.

In order to maximize gains and minimize losses, more and more researchers are involved in the practice of stock market price forecasting methods. For stock price analysis, because it has a huge amount of data and most of them are non-linear, in response to the diversity of data, numerous effective machine learning algorithms and deep learning models have been created to address the intricate relationships present in stock data. These models and algorithms have been proven over time and practice to be more efficient than traditional prediction methods (Vijh et al 2020). Most of the classical machine learning algorithms in the field today are linear regression, RWT, MACD and random forest.

Linear regression has been around for a long time and it is generally used as a model to predict quantitative responses. It is popular in machine algorithms because of its easy-to-interpret model parameters and is a very useful and widely used statistical learning method (Su et al 2012). Cakra et al. conducted a prediction of stock prices in 2015, and they concluded that the sentiment of the stockholders affects the purchasing of stocks, which leads to fluctuations in the price of stocks. Therefore, they linked sentiment analysis with stock prices and predicted the Indonesian stock market. In this study, Cakra et al. used linear regression to build a prediction model and the results of the model gave a good prediction (Cakra and Trisedya 2015). Ali et al. researchers also predicted the price of bitcoin for the next 7 days in 2020 also using linear regression model. They extracted the relevant features in the dataset with strong relation to Bitcoin price and trained the linear regression model with appropriate data chunks and ended up with a good result of



Luo, Z.
The Prediction of Google Stock Closing Price Based on Linear Regression Model and Random Forest Model.
DOI: 10.5220/0012805600004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 229-233
ISBN: 978-989-758-690-3

96.97% accuracy (Ali and Shatabda 2020). Nguyen et al. defined the authors' age prediction as a regression problem, which is a relatively new line of research. They used a linear regression model and ended up with a correlation of 0.74 (Nguyen et al 2011). Random forest models are also frequently used with prediction efforts when people use linear regression for predicting quantitative responses.

Random forests have demonstrated remarkable accuracy, which has made them a widely adopted method for numerous machine learning applications. Additionally, they are relatively straightforward to comprehend and can efficiently manage extensive datasets with high dimensionality. Nevertheless, they can be computationally demanding during the training process and might not yield optimal performance when dealing with extremely small datasets (Rigatti 2017). Kumar and Manish et al. in 2006 used both Support Vector Machines and Random Forests algorithms for stock market prediction and compared the effectiveness of the two algorithms (Kumar and Thenmozhi 2006). Mei and He et al. predicted the de facto prices in the New York electricity market by using Random Forests models and evaluated the effectiveness of the models (Mei et al 2014). Gupta et al. in 2019 used a total of five algorithmic models such as Random Forests, etc., respectively, for the diagnosed cases, dead cases and cured cases of novel coronavirus were analysed and predicted, in which random forest model outperformed the other models (Gupta et al 2021). Song et al. also predicted pressure ulcer nursing adverse event in 2022 using SVM, DT, RF and ANN respectively, and finally got the conclusion that random forest is the best performance among these four models (Song et al 2021).

The aim of this study is to predict the closing price of stocks using linear regression model and random forest model and compare the two models to find a more effective model for predicting stock prices.

# 2 METHODS

## 2.1 Data Source and Description

The historical data for the stock price of Google has been downloaded from Kaggle. The dataset includes historical data on the stock price of Google, spanning a period of 5 years from 6/11/2018 to 10/11/2023 (table 1).

To facilitate a more comprehensive observation of the information in the dataset, Table 2 show some descriptive statistical information of the dataset.

Table 1: A portion of the dataset.

| Date | Open | High | Low | Close |
|---|---|---|---|---|
| 2018/11/16 | 52.971 | 53.350 | 52.449 | 53.074 |
| 2018/11/19 | 52.860 | 53.040 | 50.813 | 51.000 |
| 2018/11/20 | 50.000 | 51.587 | 49.801 | 51.287 |
| 2018/11/21 | 51.838 | 52.428 | 51.673 | 51.880 |
| 2018/11/23 | 51.500 | 51.879 | 51.119 | 51.194 |

The objective of this study is to forecast the closing price of stocks. Therefore, the historical data trends of stock closing prices have reference significance for research. Fig. 1 demonstrates the trend of historical data on the closing price of stock.

By observing the trend of Google's stock closing price, some simple conclusions can be drawn. Firstly, it is evident that the closing price of Google's stock showed a significant upward trend over the three-year period from 2019 to 2022. Despite experiencing a period of downturn from 2022 to early 2023, it ultimately rebounded to a higher price.

Table 2: Descriptive statistical information of the dataset.

| | Open | High | Low | Close |
|---|---|---|---|---|
| count | 1254.000 | 1254.000 | 1254.000 | 1254.000 |
| mean | 96.534 | 97.666 | 95.521 | 96.614 |
| std | 30.312 | 30.599 | 29.997 | 30.285 |
| min | 48.695 | 50.176 | 48505.000 | 48.811 |
| 25% | 67.298 | 67.879 | 66.679 | 67.237 |
| 50% | 97.189 | 99.114 | 95.697 | 97.139 |
| 75% | 123.966 | 125.231 | 122.697 | 123.865 |
| max | 151.863 | 152.100 | 149.887 | 150.709 |

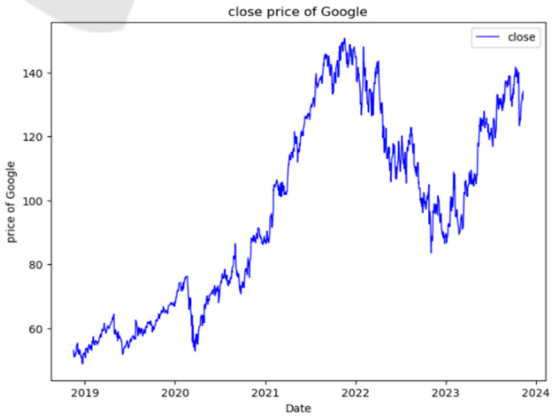


Figure 1: The trend of historical data on the closing price of stock (Picture credit: Original).

## 2.2 Variables Introduction

In order to help the model better understand the trends and patterns of stock prices and thus improve the accuracy of the predictions, this study creates seven new variables to train the model for the prediction of stock closing prices, which are the closing prices of the stock for each of the previous seven days.

## 2.3 Methods Introduction

Linear regression is an algorithm in supervised machine learning. The regression problem mainly focuses on the relationship between the dependent variable (which can be one or multiple values to be predicted) and one or more numerical independent variables (predictive variables).

Random forest regression is an ensemble learning technique that consolidates multiple decision trees to generate predictions. In this method, each decision tree within the forest is trained using a distinct, randomly chosen subset of the data. This approach aids in mitigating overfitting and enhances generalization capabilities. To obtain the ultimate prediction for a specific input, the average or weighted average of the predictions derived from all the decision trees within the forest is computed.

# 3 RESULTS AND DISCUSSION

## 3.1 Divide Dataset

During the model training process, overfitting or underfitting can be seen as an inevitable event, and whether overfitting or underfitting occurs, it can cause significant errors between the predicted results of the algorithm model and the actual results. To alleviate this phenomenon, our approach is to divide the original dataset into training and testing datasets. Since the dataset of this study uses time series data, we use TimeSeriesSplit() to divide the dataset. Table 3 presents the statistical information of the dataset utilized for both training and testing.

Table 3: Descriptive statistical information of the dataset.

| | Dataset | Training Dataset | Testing Dataset |
|---|---|---|---|
| Time Interval | 6/11/2018-10/11/2023 | 6/11/2018-17/1/2023 | 18/1/2023-10/11/2023 |

## 3.2 Forecasting Results

To evaluate the comparative effectiveness of the two models, we brought the testing dataset into the two models for testing and constructed two graphs to visualize the results predicted by the two models. Fig. 2 represents the comparison between the true stock closing price and the stock closing price predicted using the linear regression model, and Fig. 3 represents the comparison between the true stock closing price and the stock closing price predicted using the Random Forest model.

We can see that both the random forest model and the linear regression model are very good at predicting known data, and the linear regression model may be better. However, we can't accurately judge which one is better just by the graph, so we will use to specific evaluation index to compare.

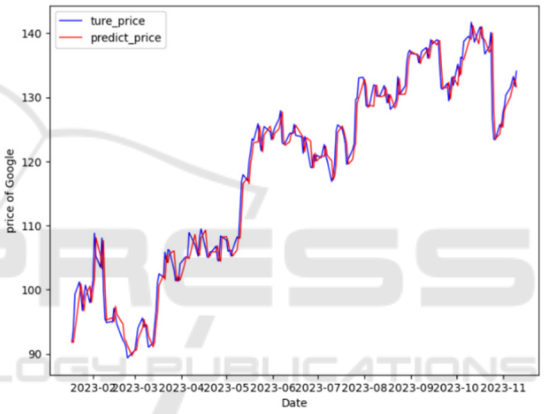


Figure 2: Predicted vs true closing stock price using Linear regression (Picture credit: Original).

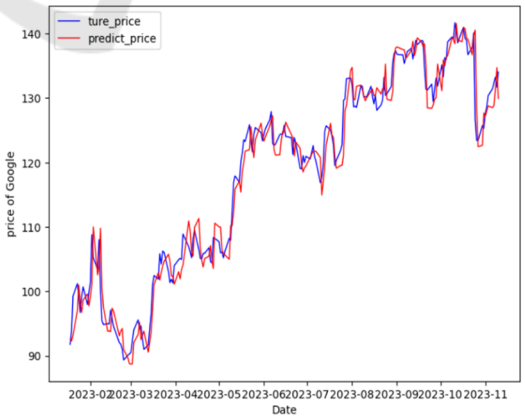


Figure 3: Predicted vs true closing stock price using Random Forest (Picture credit: Original).

### 3.3 Comparative Results

To assess the effectiveness of the models, we compare the performance of the linear regression model and the random forest model in predicting the closing price of Google Inc. We use three different evaluation metrics to measure the final minimisation error of the predicted price. Table 4 shows the results of the comparative analysis obtained using the linear regression model and the random forest model.

Table 4: Comparative analysis of the model evaluation.

| | Linear Regression | Random Forest |
|---|---|---|
| RMSE | 2.346 | 2.771 |
| MSE | 5.505 | 7.676 |
| MAE | 1.716 | 2.142 |

The result shows that the RMSE of Linear Regression is 2.346 and the RMSE of Random Forest is 2.771. The MSE of Linear Regression is 5.505 and the MSE of Random Forest is 7.676. The MAE of Linear Regression is 1.716 and the MAE of Random Forest is 2.142. Based on the evaluation metrics, it can be observed that the Linear Regression model shows better prediction results for stock prices compared to the Random Forest model. This is because the values of evaluation metrics for Linear Regression are all lower than those of Random Forest.

### 3.4 Discussion

Although it is difficult to predict the closing price of a stock, it is possible to increase the precision of the forecast and improve the forecasting efficiency with the aid of machine learning. In order to predict the closing price of Google's stock, this study took the closing price of Google's stock in the previous seven days as the independent variable and finally obtained that the closing price of Google's stock will have a stable and continuous upward trend. By comparing with the real trend of the closing price of Google's stock, it can be found that the two algorithmic models used in this study predicted the results very close to the real stock trend. The efficiency and high accuracy of the two algorithmic models demonstrate that both the linear regression model and the random forest model are efficient deep learning models that can be used to predict stock prices.

The linear regression model has a better performance than the random forest model in the comparison based on the three evaluation metrics of RMSE, MSE and MAE, which may be due to several factors. Firstly, the dataset of this experiment is small and the linear regression model is likely to converge more quickly and get better results, since it is not necessary to build a lot of decision trees; Secondly, the random forest model is an integrated learning method, but it also brings an increase in computation. Therefore, it is better to use linear regression model when the computational power is limited; in addition to that, if there is a certain linear relationship between the characteristics of the data and the target variables, linear regression usually provides more concise and easy-to-interpret results. Although the results of this study do not show that the linear regression model is a more efficient deep learning model than the random forest model, but we can know that the linear regression model may be a better choice in the above cases.

## 4 CONCLUSION

Although predicting stock prices is not an easy task, machine learning techniques have facilitated the field of stock price prediction nowadays. The aim of this paper is to test the effectiveness of machine learning algorithmic models in predicting stocks, while comparing the efficiency of two different machine learning algorithmic models.

First of all, this study uses the stock price information of Google Inc. in the past five years provided by Kaggle website, and predicts the closing price of Google Inc. stock using linear regression model and random forest model respectively, and the consequences show that both machine learning algorithmic models have good prediction results, which are very close to the real results.

In addition to this, this study also compared the efficiency of the two machine algorithm models using three evaluation metrics, and the results revealed that the linear regression model outperformed the random forest model in certain circumstances, specifically.

For future work, more machine learning algorithms can also be included at the same time for comparison in addition to the two methods mentioned above. It is believed that deeper learning of machine learning algorithmic models can lead to better results in the future in the field of stock prediction.

## REFERENCES

M. Vijh, D. Chandola, V. A. Tikkiwal, A. Kumar, Procedia comp. sci., 167, 599-606 (2020)

X. Su, X. Yan, C. L. Tsai, Wiley Interdisciplinary Reviews: Comp. Stat., 4(3), 275-294 (2012)

Y. E. Cakra, B. D. Trisedya, ICACSIS, 147-154 (2015)
M. Ali, S. Shatabda, ICAICT, 330-335 (2020)
D. Nguyen, N. A. Smith, C. Rose, ACL-HLT, 115-123 (2011)
S. J. Rigatti, J. Insur. Med., 47(1), 31-39 (2017)
M. Kumar, M. Thenmozhi, Indian inst. Cap. Mark., (2006)
J. Mei, D. He, R. Harley, T. Habetler, G. Qu, Gen. Meet. Conf. Exp., 1-5 (2014)
V. K. Gupta, A. Gupta, D. Kumar, A. Sardana, Big Data Mini. Ana., 4(2), 116-123 (2021)
J. Song, et al. Risk Mana. Heal. Pol., 1175-1187 (2021)