Stock Price Prediction of "Google" based on Machine Learning

Luna Peng[1,*]

[1]School of Art and Science, Syracuse University, Syracuse, NY 13244, US

*Corresponding author: Lpeng08@syr.edu

Abstract. By 2022, many countries have declared the epidemic's end, both an opportunity and a challenge for many investors. More and more investors are manipulating prices to influence the stock market. So investors want to predict the price of stocks to make suitable investments. The author wants to start with the platform YouTube to study the price trend of this stock and make predictions to analyze whether there are traces of the factors affecting the stock price based on linear regression and random forest regression models. The author first backtested the price of this stock and analyzed the data according to the highest and lowest day. Then, the author used the method of Linear Regression and Random Forest Regression to predict the price. The error of the Linear Regression prediction results was within 5%, within the normal range, but the Random Forest Regression 5 days prediction's accuracy is much lower (65%). It shows that the stock price prediction model--Linear Regression is more credible and is worthy of reference for investors.

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

1. Introduction

Google's stock price fluctuated wildly before and after the epidemic, and Google has paid more attention to Internet advertising in the past three years of development. Investors reduced their advertising investment in Google Internet and YouTube in the third quarter of this year. These are related to market efficiency and market efficiency. The field of financial forecasting is closely related [1].

The company's financial analysis and planning are critical to the company's future development. Rami discussed that the company's final financial situation would decline during development, but reasonable and prudent financial forecasts can at least minimize losses when force majeure occurs [1]. Financial forecasts are also critical when seeking investments, as financial forecasts will represent future financial trends and can be compared to past financial positions, providing management with valuable investment confidence. In this article, The author will focus on the price forecasting model, using two forecasting models to predict the price of Google stock and compare the accuracy, to discuss the fit between different forecasting models and companies, and the impact of forecasting of different lengths of time on the accuracy of the model [2].

In this article, the article first performs a price backtest on Google data and observes the price fluctuation range so that the prediction model can be used to discuss the next day's price. This paper used linear regression and random forest regression models to predict Google's stock price direction, using Google's price data for nearly a decade. In order to compare the two prediction models more reasonably, the author inputs the same number of variables for both models, and the training set and test set are the same. However, linear regression predicts the next day's data based on the previous ten years' data, while random regression Forest regression predicts the data for the next five consecutive days. The results show that predicting different times also affects the model accuracy.

2. Firm description

The alphabet company (GOOG) was established on 1998-09-04, providing various services, products, and platforms worldwide, striving to realize the Internet concept of a global village as soon as possible. Its main products are Google Services, Google Cloud, and other investment arm operations. Google mainly provides Internet services, including but not limited to Internet search, cloud computing, advertising technology, and other fields, and develops and provides many Internet-

based products and services. Google parent company Alphabet recently announced its financial statements for the second quarter of 2022 [3].

In the past two years, Alphabet's quarterly revenue has beaten consensus estimates eight times in a row, and only one quarterly profit has missed consensus estimates. However, under the prospect of an economic recession, Google's earnings report was not spared, and revenue and profit were worse than expected. However, thanks to the strong performance of Google's advertising business and investors' expectations for more unfavorable factors in this quarter's earnings season, Google's U.S. stock rose instead of falling after the market. It once rose more than 5% and then narrowed to 2%. Since the beginning of the year, Google has fallen by 27%, and its market value has evaporated by nearly a quarter, under-performing the 18% decline of the S&P 500 over the same period [4].

The "mainstay" advertising business has grown against the trend, and Google is still facing many challenges. As a result, analysts have slashed their estimates for Google's second-quarter earnings as they factor in the broader economic challenges advertisers now face [4].

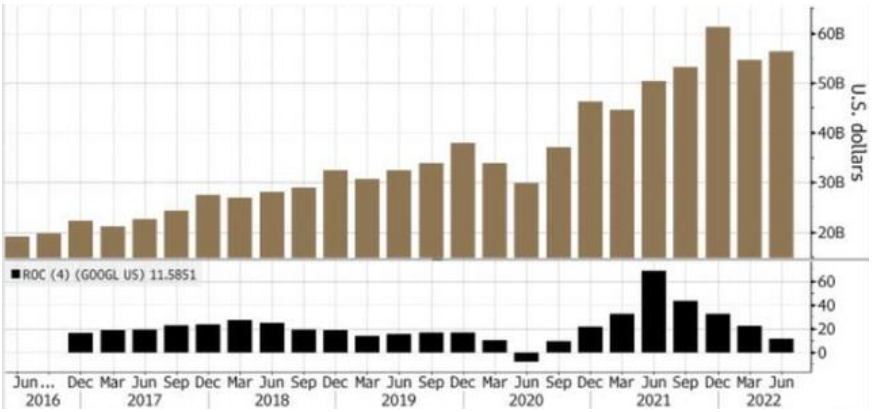

Fig. 1 Advertising Revenue

In addition, the market believes that the driving force of Google's growth will continue to be Google Cloud. However, Google's cloud business is still far behind Microsoft and Amazon, and the difference is that Google Cloud is still burning money and losing money. At the same time, its competitors operating profit margin has reached more than 20%. Google Cloud ended the quarter with a loss of $858 million. However, given the recent momentum in Google Cloud Platform and collaboration platform Google Workspace, the company has shown much higher-than-expected growth [3].

In the meantime, here are a few reasons to choose Google as a stock. First, the U.S. market index Google accounts for a significant portion of the market weight, over 7% in the Nasdaq Composite 100 and nearly 4% in the S&P 500. That said, Google's stock price is closely related to the broader market price, and we can see from the graph that Google's stock price has moved very closely with the S&P 500 over the past 12 months. So, in terms of convergence, when our price forecasting model is very accurate, it also helps with the price forecast for the S&P 500 [4].

Fig. 2 1 year price return of GOOG compared with S&P 500

The author has collected stock price data for a long time, from 2004 to 2021, which has advantages and disadvantages. However, it is also a factor in checking whether the prediction model is suitable for this stock.

3. Method and Result Analysis

3.1 General Information

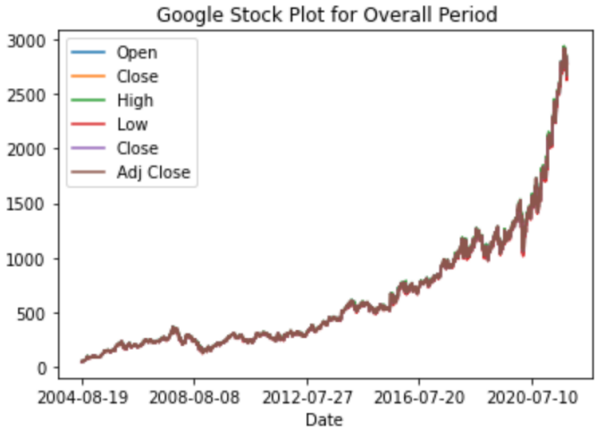

Fig. 3 Google Stock Plot for Overall Period

Fig. 3 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 [5].

Fig. 4 Google Stock Daily Returns

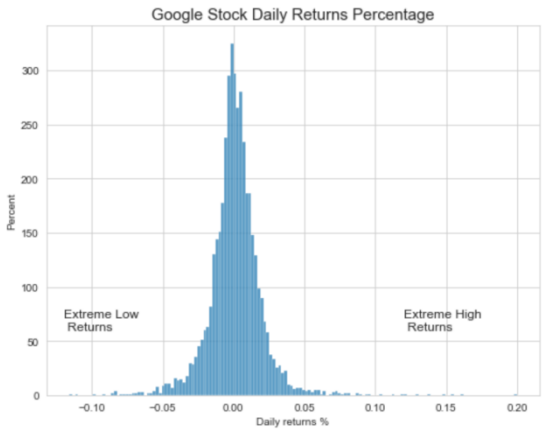

Fig. 5 Google Stock Daily Returns Percentage

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 [5].

3.2 Forecasting Model

3.2.1 Linear Regression

The relationship between two variables is a linear function - the graph is a straight line, called linear. When the variable y output is predicted as infinite and continuous, we call it regression. Linearity usually refers to the relationship between variables in equal proportions. From a graphical point of view, the shape of the variables is a straight line, and the slope is constant. The author's goal is that the value predicted by the model is infinitely close to the actual value, so the mean square error MSE is used to determine the distance. The loss function is limited to the square loss function; the ordinary least squares get an objective function. More complicated than univariate linear regression, multiple linear regression does not consist of straight lines. Analysis data: Only one independent variable belongs to the most straightforward univariate linear regression problem. A total of N+1 = 2 variables is required. The gradient descent method can solve the parameters for the parameters and bias in the independent variables. The least squares method in both linear regressions is within the framework of known data so that the total squared difference between the estimated and actual values is as slight as possible. [3, 6].

3.2.2 Linear Regression Model results

After taking data into the model, the results of next-day prediction are close to the actual data.

In Fig. 6, two lines contrast with each other; it is evident that the predicted price is always higher than the actual price. The result is the next-day prediction; the graph shows 5-day results [5].

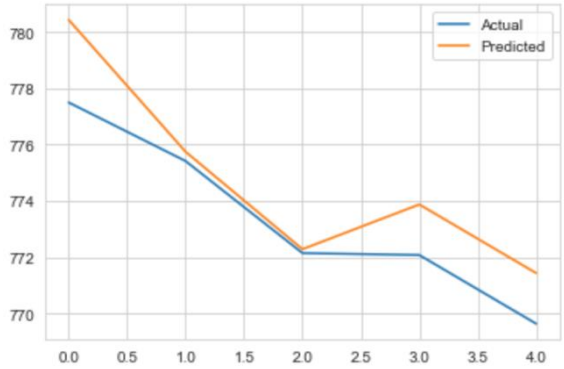

Fig. 6 Actual price vs. Predicted price

Commonly used regression model evaluation indicators:

Mean Absolute Error: 5.431318025950015

Mean Squared Error: 63.024804545909625

Root Mean Squared Error: 7.938816318942618

From the three error data above, the accuracy of linear regression prediction of next day price is high enough as the mean squared error is under 100. Moreover, the mean absolute error and root mean squared error is too small to convince the model. Errors are too small, which shows that the model data we used fluctuated too small day by day, so that is why daily return data is an essential condition to consider for the linear regression model.

3.2.3 Random Forest Regression

The random forest tree model is a particular type of bagging method. The algorithm trains several decision trees and combines their results—the average effect of each decision tree obtained by a random forest on a regression problem. The MSE is a vital regression tree indicator and can be used

as an essential indicator. The mean is usually used when using cross-validation or other means to derive results from a regression tree. Our estimate is the squared error (in the classification tree, this metric represents the accuracy of the prediction). We want the MSE to be as small as possible during the regression. However, the regression tree interface point returns the square of R, not that of the MSE. Usually, a simple average method is used for regression. The average value of the regression results obtained by the disabled learner T is the final output of the model. [6, 7].

When the author enters open, high, low, close, and volume as variables, this will allow each factor to impact the final prediction.

3.2.4 Random Forest Regression Model Results

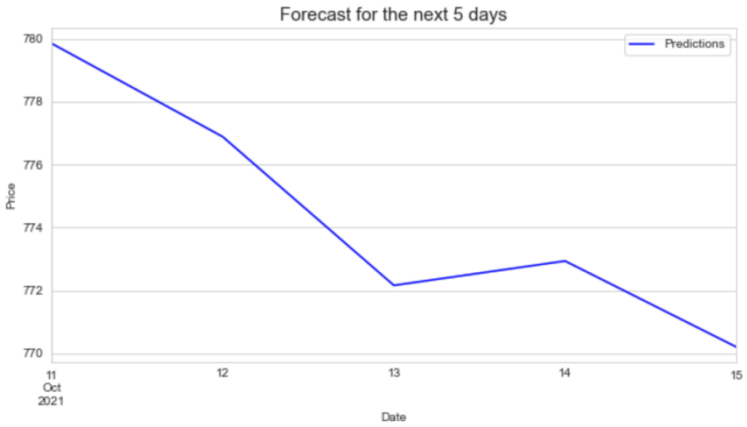

Fig. 7 Forecast for the next 5 days

Fig. 7 draws the line for the next five days after the last day in data, but in the real world, the prices of these days are different, which means these points are out of the line.

As the daily returns graph draws that the everyday price floating is not violent and the price around Oct. 2021 is nearly 2800+, the forecasting price is just 700+ [5].

Mean Absolute Error: 547.5078
Mean Squared Error: 567592.6865
Root Mean Squared Error: 753.3875
(R^2) Score: -1.1122
Accuracy: 65.92 %.

Even though using the same variables as inputs, the errors are 100 times higher than linear regression for Random Forest Regression. R2 number is negative only when the chosen model does not follow the trend of the data. It can be a case of over-fitting in the model [7, 8]. Nevertheless, it also shows that five-days forecasting accuracy is much less than the next-day forecasting of linear regression.

4. Discussion

4.1 Linear Regression

The linear regression model is a simple model that does not require particularly complex operations, and it can be performed quickly even with a large amount of data. With coefficients, individual variables can be understood and explained. However, linear regression requires strict assumptions and is very sensitive to anomalies and differences in the input data. If different features are not independent, it may lead to collinearity between components, resulting in inaccurate models. Linear regression also has problems such as collinearity, autocorrelation, and heteroscedasticity [8, 9].

4.2 Random Forest Regression

The dataset of random forests is fine. The introduction of quadratic random variables makes random trees less prone to overfitting, but there is still the possibility of overfitting. The introduction of the quadratic random feature makes the random tree have better anti-noise performance; it can be well adapted to the data set without the need for feature selection. Furthermore, random forest regression is very fast to train and can classify variables according to their importance; interactions between features can be discovered during training. In this way, the regression of random forests is easy to parallelize and easy to implement [6, 7].

Although the random forest regression model is easy to operate, it may not work well for small datasets and low-dimensional data. Moreover, the entire model is a black box with no reliable explanatory power. Finally, the two randomnesses of random forest regression can lead to inconsistent results [10-15].

5. 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.

References

[1] Rami, C. (2022, September 14). Could google trade lower? exploring the bear case arguments. Retrieved September 17, 2022, from https://seekingalpha.com/article/4541035-could-google-trade-lower-exploring-bear-case

[2] Ali, R. (2020, October 15). The Why Behind Financial Forecasting. Retrieved September 16, 2022, from https://www.netsuite.com/portal/resource/articles/financial-management/importance-financial-forecasting.shtml

[3] King, T. (2020). Detailed in Linear Regression model. Retrieved September 17, 2022, from https://blog.csdn.net/iqdutao/article/details/109402570

[4] Krause, R. (2022, September 06). Is Google a buy or sell as investors mull growth beyond digital ad business? Retrieved September 9, 2022, from https://www.investors.com/news/technology/google-stock-buy-now/

[5] Abhi. (2021, October 12). Google stock price (all time). Retrieved September 9, 2022, from https://www.kaggle.com/datasets/akpmpr/google-stock-price-all-time?resource=download

[6] Y. Lin, "Research on Business Development Strategy of Application Software Store," IEEE Xplore, Oct-2020. [Online]. Available: https://ieeexplore.ieee.org/document/9607174.

[7] Dai, L. (2021). Random Forest in Machine Learning. Retrieved September 17, 2022, from https://blog.csdn.net/lindaicoding/article/details/119388694

[8] Brownlee, J. (2020, August 14). Linear regression for machine learning. Retrieved September 9, 2022, from https://machinelearningmastery.com/linear-regression-for-machine-learning/

[9] Amolambkar. (2021, January 16). Stock price prediction using linear regression. Retrieved September 9, 2022, from https://www.kaggle.com/code/amolambkar/stock-price-prediction-using-linear-regression

[10] Vanshikhaangrish. (2022, March 20). Netflix stock prediction- Random Forest & Linear R. Retrieved September 9, 2022, from https://www.kaggle.com/code/vanshikhaangrish/netflix-stock-prediction-random-forest-linear-r/data

[11] Support Vector Machines. (n.d.). Retrieved September 9, 2022, from https://scikit-learn.org/stable/modules/svm.html#classification

[12] Anderson, O. D., & Perryman, M. R. (1981). Time Series analysis. Amsterdam: North-Holland Pub.

[13] Beaver, W. H. (2002). Perspectives on recent Capital Market Research. The Accounting Review, 77(2), 453-474. doi:10.2308/accr.2002.77.2.453

[14] Peixeiro, M. (2022, September 06). The Complete Guide to Time Series Analysis and forecasting. Retrieved September 9, 2022, from https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

[15] Time Series Forecasting methods. (2022, May 04). Retrieved September 9, 2022, from https://www.influxdata.com/time-series-forecasting-methods/