Stock Prediction of Google based on ARIMA, XGBoost and LSTM

Junchen Yao*

Department of mathematics, Nanjing University of Information, Science and Technology, Nanjing, China

*Corresponding author: fl805082@student.reading.ac.uk

Abstract. In the recent years, google has become one of the most powerful companies in the world, due to its big market dominance. More and more people want to predict the stock price of google, however changes in the stock price are hard to find because they combine with social and economic development. Therefore, many different models which can be divided into traditional-based model, machine learning and deep learning models are designed to improve the accuracy of stock price prediction. This paper firstly compared three high-frequency used different models based on different aspects: autoregressive integrated moving average (ARIMA) model, eXtreme Gradient Boosting (XGBOOST) model and Long short-term memory (LSTM) model. mean absolute error (MAE), mean squared error (MSE), rooted mean squared error (RMSE), r-squared(R2) are presented due to the performance of models. Empirical results show that XGboost model provide more accurate approximation than ARIMA and LSTM models. In addition, the accuracy of LSTM is the worst.

Keywords: Google; ARIMA; XGBOOST; LSTM.

1. Introduction

stock prediction has been a big issue when the financial market first appears. Due to the capricious and inconstant financial market, predictive modelling has changing among the times. The main predictive models can be divided into two basic categories: statistical techniques and soft computing techniques. [1]. Google is a business company that focuses on artificial intelligence, cloud computing, computer software, search engine technology, and internet advertising. Due to its significant market dominance and reputation as "the most powerful company in the world," it has also been referred to as one of the most valuable brands in the world. While The 20th century is an age of information explosion, personal data becomes the most valuable information. While Google is the most competitive search engine [2], it is natural that it become a big data collector. In the case of concerning the perform of this big data collector in the stock market, the paper found the data in Kaggle and tries to analysis the stock price using three different model: ARIMA, XGBOOST, LSTM.

Before the flourishing of AI techniques, autoregressive integrated moving average (ARIMA) is a high-frequency used statistical techniques for making predictions using past observations and it is an extension of an ARMA model, which stands for autoregressive moving average. To better comprehend the data or to forecast upcoming series points, both of these models are fitted to time series data (forecasting). While ARIMA is mostly solving stationary data, and shows a high accuracy in making predictions, it has limitations for example non-stationarity, seasonality, and other factors [3]. ARIMA model has been called the traditional-based algorithms compared with the brand-new learning-based time series data forecasting algorithms for example “Long Short-Term Memory (LSTM)” [4], but the perform of ARIMA did not mean to be lower than the learning-based algorithms which will proved in many comparative studies in the next section.

During the recent decades, the appearance of big data and the speedy increase of artificial intelligence (AI) techniques, machine learning (ML) and deep learning (DL) skills has been widely used in the stock prediction area in the stock market [5]. The skills that enable computer systems to improve their behavior for a given task by learning from previous experiences are known as machine learning. [6]. ML includes linear regression, logistic regression, decision tree, SVM algorithm and so on. D Neural network architectures or sets of labeled data with multiple layers are used to train deep learning models. DL includes Convolutional Neural Networks (CNNs), Long Short-Term Memory

Networks (LSTMs), Recurrent Neural Networks (RNNs) and so on. The paper chooses the model that the most representative model in the areas of machine learning and deep learning.

XGBoost is a decision tree promotion model proposed by [7], containing a strong classifier integrating several tree models [8]. XGBoost can solve both classification and regression problem which using several given features its use in the stock market is nearly nonexistent [9]. To compare with ARIMA, XGBoost model is better to deal with the features of non-stationary data and make predictions about these factors. However, a long time may be taken due to the large dataset [10].

Long short-term memory (LSTM) is one of the most important and widely used method of artificial neural networks [11]. As the difference to ANN model, LSTM has a memory unit that are designed specially to deal with the problem of vanishing gradients and exploding gradients that can be encountered when training traditional RNNs [12]. The network is able to effectively allocate memories and remotely enter them at any time thanks to these memory cells, thereby dynamically and highly predictively capturing the data structure over time.

The paper employs three fundamental methods: ARIMA, XGBoost, and LSTM for stock price prediction, which will be discussed in the following section. However, the paper did not make further predictions in the concrete days. The main purpose for the paper is to compare the existed model and help to understand the accuracy of the model in the daily financial market. Therefore, in the paper, we construct model and test the model to get the accuracy. The whole coding processing is finished by Python. The literature review and associated works are described in Section 2. The general ideas for each of the methods are presented in Section 3, after which the models customized for Google Stock are created. The accuracy with error measurements is used in Section 4 to compare the outcomes from each of the three models. Section 5 contains the conclusions.

2. Methodology

The methodology section contains five parts. The first part describes the dataset that are used in the paper. Then, in the following parts describe the theoretical basic of each model and decomposition function, and focus on the problem that how can the models fit in the dataset.

2.1 Data

Google stock data is found in kaggle.com from 19 August 2004 to 24 March 2022. Six columns make up the dataset: Date, Open, Close, High, Low, and Volume. It is a daily dataset and close price is the main aspect for analysis. Other index can be seen as the features of XGBoost model. Figure 1 shows the close price of google stock.

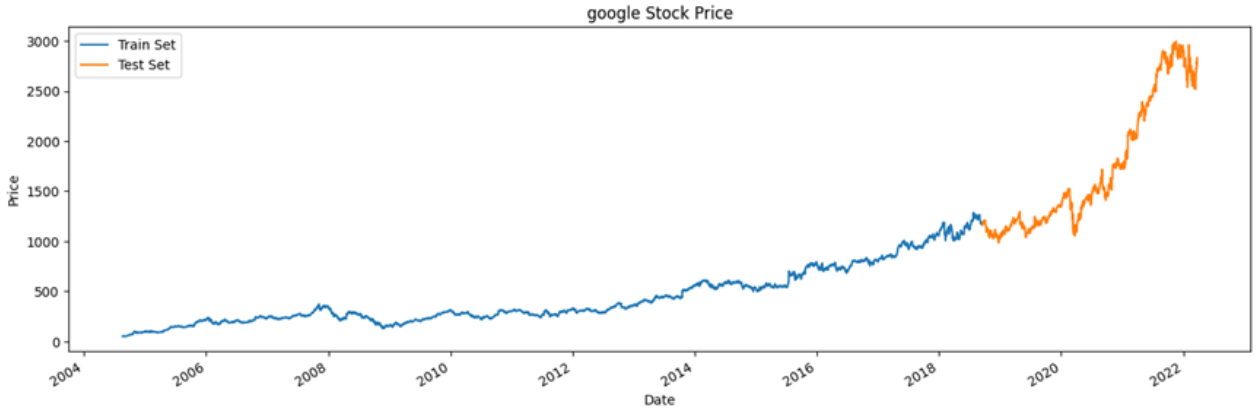

Figure 1. The google stock price

Form figure 1, we can see that train set and test set is divided into 8:2. The train test is the dataset that train the models and test set is for the prediction and error measures of the trained model.

2.2 Autoregressive Integrated Moving Average Model

2.2.1 ARIMA (p, d, q) models

Stationary is a main requirement for the time-series analysis. However, nature data in the finance market are normally non-stationary which need some parameters in the model to amend the dataset into a smooth curve. Using ARIMA (p, d, q) to fit in the data where the autocorrelation function determines both the moving average, MA(q), order q, and the autoregression AR(p), order p; the degree of differentiation that results in a stationary ARMA (p, q) process is referred as d.

2.2.2 ARIMA (p, d, q) model for Google stock

Figure 1 shows that Google stock is non-stationary data where it shows an upward trend. Due to the operating system is python, it gives an option that computer can automatically select the fit parameters and AIC. Table 1 gives the results that fit for the data.

Table 1. AICs of Giving Parameters in ARIMA Model

ARIMA(p, d, q)	AIC
(2,1,2)	24403.054
(0,1,0)	24403.465
(1,1,0)	24397.250
(0,1,1)	24397.215
(0,1,0)	24407.721
(1,1,1)	24399.059
(0,1,2)	24399.196
(1,1,2)	24401.197

So (0,1,1) becomes the best parameters due to the minimization of AIC.

2.3 Extreme Gradient Boosting Model Processing

This approach, which belongs to the family of gradient boosting techniques, is very scalable. When compared to the current tree-based algorithms, it has a much faster learning rate and higher prediction accuracy since it achieves parallelization and decentralization. The main different between XGBoost and GBDT is the definition on object function where XGBoost use Taylor expansion to approximate the object function. In addition, XGBoost can do classification problem and regression problem simultaneously.

Using the existed index in the dataset, features selection divided the features into four parts: close, range of open and close, range and high and low and volume. Because we did not know the exact information in the same day of the close price, one to three days before the day can be features to the XGBoost model. 12 features are made to be added into XGBoost in total.

2.4 Long Short-Term Memory Model Processing

2.4.1 Long short-term memory

LSTM is an artificial neural network (ANN) containing a special cell used for save long-term memory and short-term memory. The main problem that LSTM solved is the gradient vanish (tend to zero) and the gradient explode (tend to infinity) happened in ANN. There are three main stages that LSTM unit perform. Firstly, the forgetting stage, Selective forget the import data, mainly forgetting the unimportant features and keep the important ones. Then selective memory stage memory the important features. Finally, output the data. Figure 2 shows the basic function of the LSTM unit. forget and input gates are containing inside the figure 2.

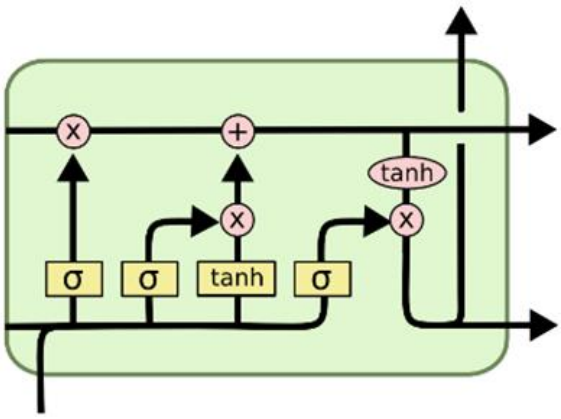

Figure 2. LSTM unit

There are more basic in ANN that contains input, hidden and output layers. These layers form a network together, using fully connected, which meaning all nodes need to be connected. In addition, the activation functions play an important part in the data processing because it transforms the data into the ones that fit in the model. The sigmoid and tanh activation function are frequently used in the neural network, which are shown in equations (1) and (2).

$$S(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$tan^{-1}(x) = \frac{1-e^{-x}}{1+e^{-x}} \tag{2}$$

2.4.2 Long short-term memory for google stock

In the LSTM model for google stock, the input layer consists one main consideration: close price. As the network structure, 1-1-1 is the main consideration because if the number of hidden layers is too large, the model is easily overfitting.

3. Results and Discussion

In the section 4, the results perform in the chart that the predicted value in the test set compared with the real value. The overall performance of each model can be checked by four different error measures, respectively Mean Absolute Error (MAE), Mean Squared Error (MSE), Rooted Mean Squared Error (RMSE), R-squared(R^2). The formulas (3), (4), (5), (6) to calculate these errors are as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - f(x_i)| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2} \tag{5}$$

$$R - Squared = \frac{SS_{regression}}{SS_{total}} \tag{6}$$

y_i are the true values, $f(x_i)$ is the predicted value, $SS_{regression}$ is the sum of squares due to regression and SS_{total} is the total sum of squares

3.1 ARIMA Model Results

Using ARIMA (0,1,1), we train our ARIMA model and get the performance via the figure that compared predicted value to the real value, and also the analysis of four error measures in the Equation (1), (2), (3), (4). The results are showed in the figure 3 and table 2.

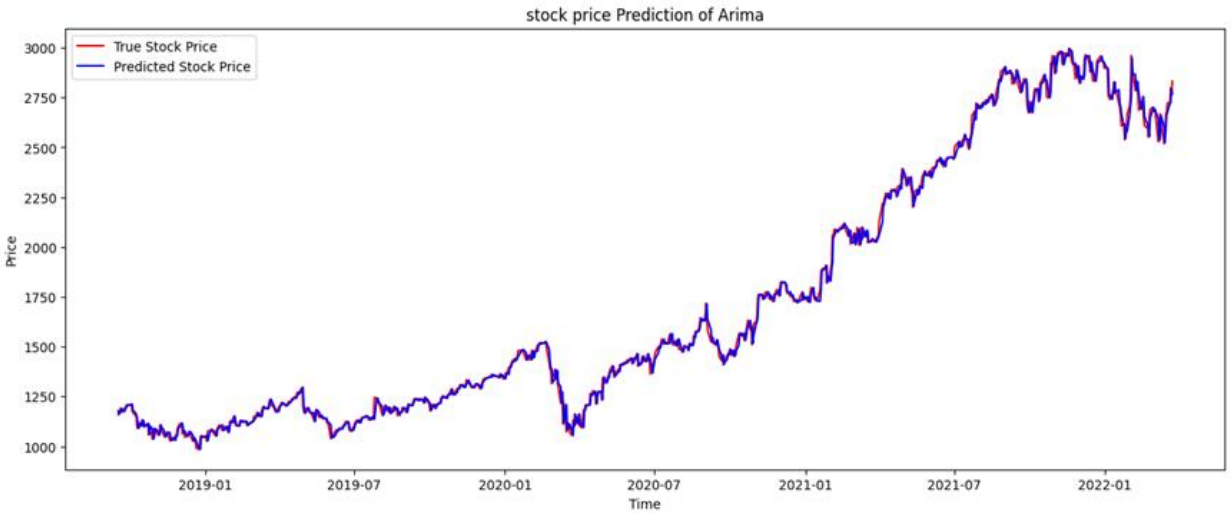

Figure 3. Stock Price Prediction of ARIMA

Table 2. Results of ARIMA

Model	MAE	MSE	RMSE	R-squared
ARIMA	24.68	1254.69	35.42	0.9969

Figure 3 presents a perfect matching for the predicted value. In addition, table 2 shows a lower MAE, MSE, RMSE and a higher R-squared compared with ARIMA. We cannot totally deduce that XGBoost has more excellent performance than ARIMA, for XGBoost has more input and more results of features importance. Combining table 3 and 4, we can make our conclusion that close lag 1 which is also known as yesterday close price data dominate the prediction, for its importance is 0.9868. Therefore, if there have only close price of one feature, the results will be similar to the previous results. Concerning the most important feature dominant the prediction, we say that XGBoost model fit better than ARIMA model do.

3.2 XGBOOST Model Results

As we construct 12 features, the results conclude the importance of these features, the features importance which are showed in table 5 and figure 4 shows the predicted value compared with the real value. In addition, the error measures are showed in the table 4.

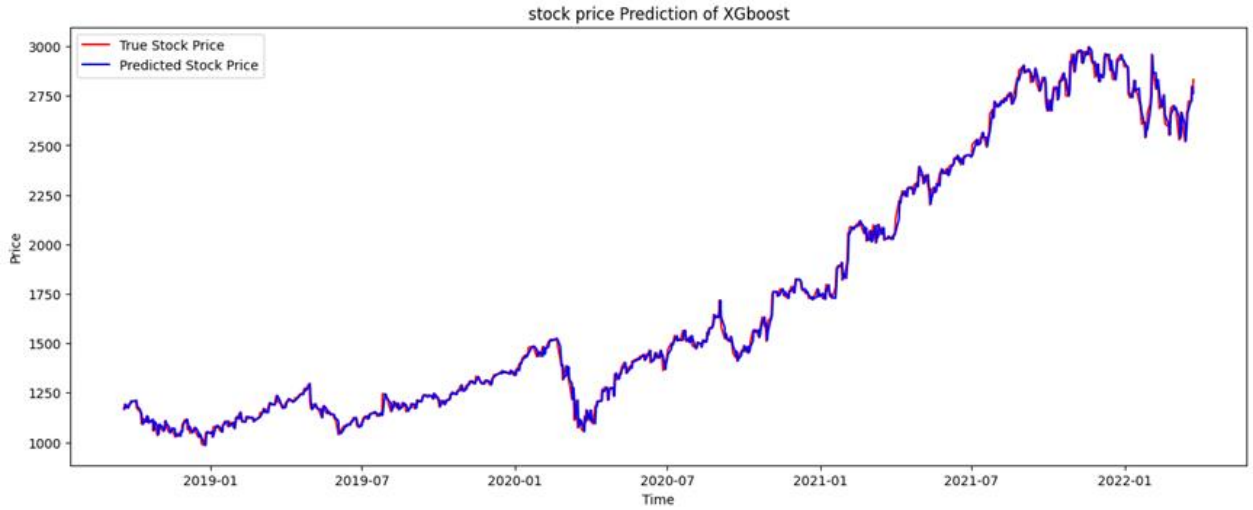

Figure 4. Stock Price of XGBoost

Table 3. Results of ARIMA

Model	MAE	MSE	RMSE	R-squared
XGBoost	22.06	1048.70	32.38	0.9974

Figure 4 presents a perfect matching for the predicted value. In addition, table 4 shows a lower MAE, MSE, RMSE and a higher R-squared compared with ARIMA. We cannot totally deduce that XGBoost has more excellent performance than ARIMA, for XGBoost has more input and more results of features importance. Combining table 4, we can make our conclusion that close lag 1 which is also known as yesterday close price data dominate the prediction, for its importance is 0.9868. Therefore, if there have only close price of one feature, the results will be similar to the previous results. Concerning the most important feature dominant the prediction, we say that XGBoost model fit better than ARIMA model do.

Table 4. Features importance

	Close lag 1	Close lag 2	Close lag 3	Range hl lag 3
Importance	0.9868	0.0115	0.0010	8.527e-05

3.3 LSTM Model Results

Figure 3 shows the stock price prediction of LSTM, and the errors measures in the table 5. The results show mismatching between the prediction value and the real value of google stock. Figure 5 shows the huge mismatching roughly started at the middle of the test set and deviation became bigger and bigger. The final results are also showed in a bad performance, as the RMSE is even four times the one in XGBoost. All index of prediction is extraordinarily higher than those in ARIMA and XGBoost, which means exactly the LSTM of 1-1-1 is not fit in the google stock price.

Figure 5. Stock Price Prediction of LSTM

Table 5. Results of LSTM

Model	MAE	MSE	RMSE	R-squared
LSTM	80.12	14895.97	122.04	0.9633

The results show mismatching between the prediction value and the real value of google stock. Figure 5 shows the huge mismatching roughly started at the middle of the test set and deviation became bigger and bigger. The final results are also showed in a bad performance, as the RMSE is even four times the one in XGBoost. All index of prediction is extraordinarily higher than those in ARIMA and XGBoost, which means exactly the LSTM of 1-1-1 is not fit in the google stock price.

3.4 Comparison of Results of Three Models

Table 6 presents the whole error measures; it is clear that ARIMA and XGBoost model perform better than LSTM.

Table 6. Table of results comparisons

Model	MAE	MSE	RMSE	R-squared
ARIMA	24.68	1254.69	35.42	0.9969
XGBoost	22.06	1048.70	32.38	0.9974
LSTM	80.12	14895.97	122.04	0.9633

We discovered that the LSTM model's predicting accuracy level is not quite noteworthy in comparison to the other two models based on the empirical findings shown in table 6. Constructing neural network is a complex process where experience is needed to fit for the realistic question. Therefore, a bad network constructing may be a reason for the bad performance of LSTM. The overall performance of ARIMA and XGBoost proved these two models can be applied to more complex question.

4. Conclusion

This study compares the performance of three models—ARIMA, XGBoost, and LSTM—to forecast Google stock prices. The results are matched to those articles that shows ARIMA models perform better over ANN models. in addition, as we add XGBoost into our comparative study, we can do a deeper analysis that one hidden layer LSTM has some disadvantages compared with the XGBoost and ARIMA model. One reason maybe the LSTM is too simple for the non-stationary data.

Our results are in contradiction to the articles which shows a high accuracy in the ANN models compared with ARIMA.

In the future study, for the LSTM model, there can be more hybridization of existing model, or some decompositions perform on the original data. As for XGBoost model, more index should be involved and participated in the features engineering. Therefore, more input can be put into XGBoost model or LSTM model to get more accuracy of the models. For ARIMA model, some filter or smoothing method can be done in the data processing section, to make the data more stationary for ARIMA to fit in.

References

[1] Adebiyi A A, Adewumi A O, Ayo C K. Comparison of ARIMA and artificial neural networks models for stock price prediction. Journal of Applied Mathematics, 2014.

[2] Giustini D. How Google is changing medicine. BMJ, 2005, 331 (7531), 1487 - 1488.

[3] Siami-Namini S, Tavakoli N, Namin A S. A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE international conference on machine learning and applications (ICMLA) (pp. 1394-1401). IEEE. 2018.

[4] Yun K K, Yoon S W, Won D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. Expert Systems with Applications, 2021, 186, 115716.

[5] Chollet F. Deep learning with Python. Simon and Schuster. 2021.

[6] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, 785 - 794.

[7] Yang Y, Wu Y, Wang P, Jiali X. Stock price prediction based on xgboost and lightgbm. In E3S Web of Conferences (Vol. 275, p. 01040). EDP Sciences. 2021.

[8] Nobre J, Neves R F. Combining principal component analysis, discrete wavelet transforms and XGBoost to trade in the financial markets. Expert Systems with Applications,2019, 125, 181 - 194.

[9] Alim M, Ye G H, Guan P, Huang D S, Zhou B S, Wu W. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. BMJ open, 2020, 10 (12), e039676.

[10] Gao T, Chai Y, Liu Y. Applying long short term momory neural networks for predicting stock closing price. In 2017 8th IEEE international conference on software engineering and service science (ICSESS) IEEE. 2017, 575 - 578.

[11] Cheng L C, Huang Y H, Wu M E, Applied attention-based LSTM neural networks in stock prediction. 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4716 - 4718, doi: 10.1109/BigData.2018.8622541.

[12] Roondiwala M, Patel H, Varma S. Predicting stock prices using LSTM. International Journal of Science and Research (IJSR), 2017, 6 (4), 1754 - 1756.