Predicting BMW Stock Price Based on Linear Regression, LSTM, and Random Forest Regression

Sijian Tao[1, *]

[1] College of Literature, Science, and Arts, University of Michigan, Ann Arbor, Michigan, the United States

*Corresponding author: taosa@umich.edu

Abstract. The stock market is one of the most significant aspects of a county's economy. Stock price prediction is not easy to implement due numerous features influencing it. This study will choose BMW as the target company, and predict its stock price with various state-of-art scenarios. The central claim of this article is to construct three reliable models that could predict the stock price of BMW by extracting and analyzing previous days' stock prices. The major analysis is done by modeling and exploring data analysis, and is composed of several graphical methods. For the modeling part, three models (i.e., Multiple Linear Regression, LSTM, and Random Forest Regression) are constructed to predict BMW stock price. The database consists of five years of BMW stock prices on Kaggle to explore more analysis. Then, step by step checking and demonstrating are presented to prove the feasibility and precision of these three models. According to the analysis, the most precise model that has the highest accuracy and lowest error by analyzing regression and model results. These model final outcomes indicate that the Multiple Linear Regression model presents a higher accuracy and lower mean squared error than LSTM and Random Forest Regression. These results shed light on guiding further exploration of BMW stock price prediction.

Keywords: Stock price prediction; LSTM; Linear Regression; Random Forest; Machine Learning.

1. Introduction

In the past few decades, stock price prediction become popular globally and gain more and more attention from investors in the stock market around the world. Investors are interested in predicting stock because an accurate stock price prediction could bring enough profit to investors [1-4]. However, stock price prediction is not easy to implement, it contains many features that could affect the price. Many past research papers focused on the stock market price prediction using different models [4-8]. Likewise, this study will use the idea of the linear regression model to find the best fit, using long short term memory, and random forest regression. This paper will predict the stock price of BMW company.

BMW used to rank first place in top-selling luxury cars globally; however, it lost its top place to Mercedes-Benz in 2015 [9]. Plenty of investors used to consider BMW as their first choice of traditional luxury cars. However, with the risen of new energy automobiles, like Tesla, and with the development of the innovative powerful automobile cars from other brands (e.g., Mercedes-Benz, Audi, and Lexus), BMW's sales went down. It lost the crown. BMW's brand has become stale and boring is the reason that BMW lost first place in luxury selling cars until now [9, 10]. Due to too many consensus driven, BMW needs to innovate more. Thus, they hire a new CEO in 2019, which aims to reinvigorate the brand until now [10]. By focusing on the changing in BMW, its stock price is attracting by seeing the fluctuation of the stock price from 2016 to 2021. This research is dedicated to using linear regression, long short term memory, and random forest regression to compare which model predicts the most accurate BMW stock price.

In efforts to search for older studies that could help collect information and provide insights into analysis, discussing the following relevant papers that helped organizing the topic step by step will be the primary section of this paper. Before moving to the prediction part of the paper, learning some stock market analysis based on fundamental and technical analysis is significant. Thus, researchers state that the basic idea of the stock market and the reason why investors want to predict stock price.

The researcher gives the basic frame of machine learning and an introduction to the LSTM model, which give a brief understanding of the LSTM model step by step [4].

Moreover, moving on to the machine learning algorithm, the difference between the traditional approach to stock price and the machine learning approach is crucial. Other scholars argue that the concrete comparison result between traditional statistical method and machine learning method. The researcher states a detailed methodology in this article, introducing a fundamental thought about the linear regression and random forest model [5].

Besides learning linear regression and random forest, LSTM modeling is another exciting algorithm. By understanding the fundamental theory of LSTM, there is an article that introduces the stock price prediction based on LSTM. Other researchers also discuss the RNN and states that the shortage of RNN and how long short term memory satisfies the long-term dependence. Researchers use a two-year period of data from 2016 to 2018 to predict the stock price of TCS company. They apply two algorithms to make the prediction. The first algorithm is RNN, and the second algorithm is LSTM [6].

Furthermore, doing deep learning and research on the Random Forest algorithm is another necessary method. Previous literature states that the Random Forest algorithm in the methodology section concretely, and the researchers compared the random forest model with ANN, which means that this article indicates more comparison between ANN and random forest. This article helps building a strong understanding of the random forest [7].

Based on above literature analysis, this paper will introduce the basic algorithm for the following models: Linear regression, LSTM, and Random Forest and analyze the model conclusion. The rest part of the paper is organized as follows, The Sec. 2 will indicate the dataset and model preparation. The Sec. 3 will visualize the result and give a comprehensive conclusion based on comparison. Subsequently, the Sec. 4 will demonstrate the limitations and future outlooks of this study. Eventually, a brief summary will be given in Sec. 5.

2. Methodology

2.1 Data & Variables

The historical five years data is downloaded from Kaggle. The dataset includes from the year 2016 to 2021 BMW's stock price for each individual day. There are seven column in this dataset, which are the date, open price, high price, low price, close price, adjusted close price, and volumn. Since the aim of this article is to predict the close price of BMW's stoc, the close price trend is illustrated in Fig. 1.

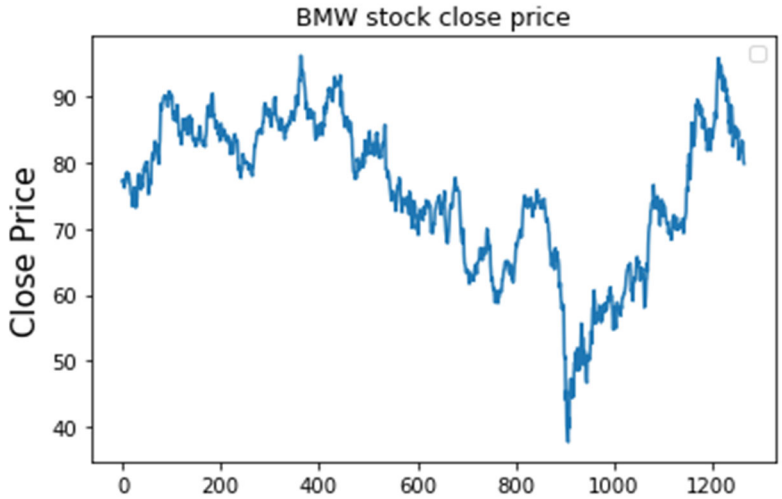

Fig. 1 The close price of BMW.

2.2 Correlation analysis

After reading the data into python notebook, the importance of each variable that related to the 'Close price' is crucial. Therefore, carrying out a correlation analysis before building models is necessary. Seen from the Table. 1, volumn presents the weakest correlation degree with close price. Thus, this paper will drop the volumn column from the dataset and ignore this variable later in three models.

Table. 1 Correlation Analysis Table.

	Open	High	Low	Close	Adj Close	Volumn
Open	1.000000	0.998250	0.998247	0.996543	0.862615	-0.277303
High	0.998250	1.000000	0.998115	0.998614	0.866445	-0.260020
Low	0.998247	0.998115	1.000000	0.998527	0.863146	-0.295707
Close	0.996543	0.998614	0.998527	1.000000	0.866205	-0.277946
Adj Close	0.862615	0.866445	0.863146	0.866205	1.000000	-0.330797
Volumn	-0.277303	-0.260020	-0.295707	-0.277946	-0.330797	1.000000

2.3 Multiple Linear Regression

Linear regression is a model that could assume there is a linear relationship between the dependent variable and independent variable. The independent variable is the input variable, and the dependent variable is the single output variable. A single linear regression means that there is only one input variable. When there are multiple input variables, the corresponded method called multiple linear regression [9]. This paper uses the multiple linear regression model to make prediction. The fomula is given as follows:

$$Y = \beta_0 + \sum_{i=1}^{5} \beta_i X_i \quad (1)$$

where Y is the dependent variable, which is the today's close value that one wants to predict, X_i is the independent variable, which is the inout features that this paper selected for model training. They are today's open price, yesterday's high price, yesterday's low price, yesterday's close price, and yesterday's adjusted close price. The main idea is to use yesterday's stock price dataset plus today's open price to predict today's close price.

2.4 Long Short Term Memory

Long Short Term Memory, known as LSTM, which is a kind of RNN that originally designed to enable networks to learn long-term dependence in time series data. LSTM has four major components, which are forget gate, input gate, output gate, and cell state.

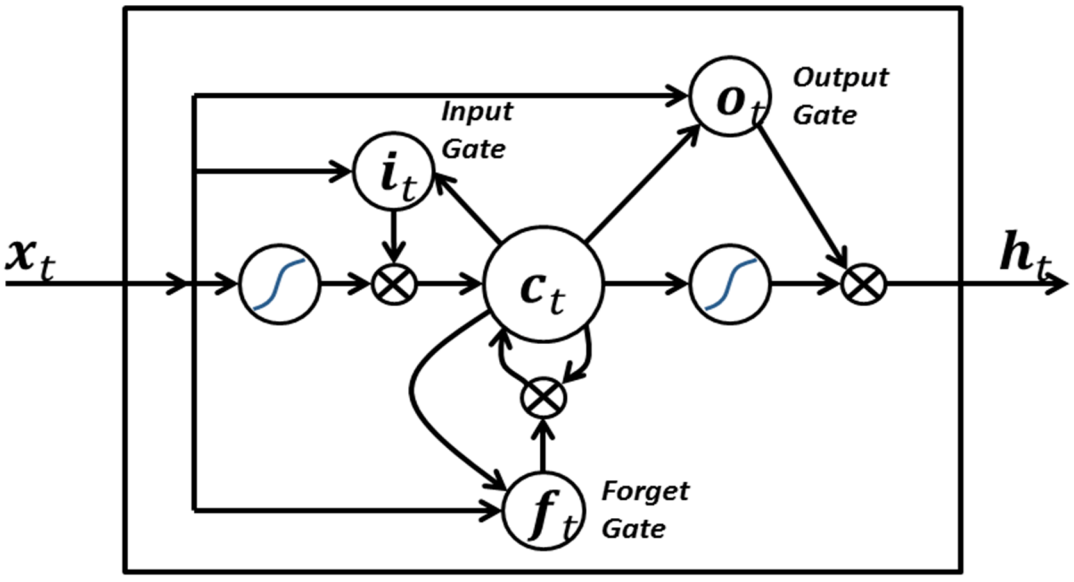

Fig. 2 A skecth of the LSTM [8].

As shown in Fig. 2, it is the regular LSTM model operating method. The formulae are as follows:

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \quad (2)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (4)$$

$$\widetilde{c_t} = tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_f\right) \quad (5)$$

$$c_t = f_t c_{t-1} + i_t \widetilde{c_t} \quad (6)$$

$$h_t = o_t \tanh(c_t) \quad (7)$$

Here, f_t stands for forget gate, which is a sigmiod function that controls the total information from the previous cell state and the current input that will flow into the current cell state; i_t stands for input gate that controls the amount of information flowing from the current input to the cell state; o_t stands for output gate, such that controls how much information goes into the hidden state from the current cell state; c_t stands for cell state that storage unit for long-term memory; h_t stands for hidden state that storage unit for short-term memory. By operating the LSTM algorithm in my project, this paper use 30 previous days' close price value to predict the next day's close price.

2.5 Random Forest

Random Forest is a machine learning technic. Random forest consist both classification and regression algorithm. In addition, random forest is an ensemble algorithm consisting of decision trees. There are four steps to establish a valid random forest. Firstly, for a sample that has 'N' sample size, it is taking 'N' times to put it back, one at a time, and finally 'N' samples are formed. The 'N' samples that are choosed are used to train a decision tree, and these samples will act as the root node of the decision tree. Secondly, when the nodes in the decision tree supposed to be splited, and every individual sample has totally 'M' attributes, 'm' attributes will be selected from the total 'M' attributes randomly, which would satisfy the condition that m < M. Then, one needs to select one attribute as the splitting arrtibut from these 'm' attributes, for instance, information gain. Thirdly, during the formation of the decision tree, each node should split according to step 2 until it can not split anymore. Finally, a large number of decision trees can be established by following step 1 to 3, so as to form a random forest.

3. Results & Discussion

. To evaluate which model present the best prediction on BMW stock price, this paper makes a comparison between those three models. By comparing the model conclusion, this article uses Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared thoese three technical standard to make a conclusion. Root Mean Square Error (RMSE) is the standard deviation of the residuals, which is the prediction errors. Moreover, RMSE is a measure that could test how spread out residuals are. In other words, it could tell you the concentration of data that around the prediction line. R-sqaure is also known as the coefficient of determination, R means that the prediction model explains the proportion of variance of variables. Variance measures the dispersion or fluctuation range of variable values. The smaller the variance, the smaller the fluctuation of variable values. In other words, the easier the value of variables is to be predicted and guessed. Assuming R-square =0.8, it means that after fitting, the variance of the variable is reduced by 80%, and the value of the variable is easier to be determined. MAE is actually the error average in a more general form. If an error is [-1,0,1], then the average value of the system is 0, but this does not mean that there is no error in it. We need to add the absolute value to confirm that there is no error. Therefore the formula is:

$$MAE = \frac{1}{n}\sum_{k=1}^{n} |actual_k - predict_k| \quad (8)$$

where $actual_k$ is the actual close price value, and $predict_k$ is the predict close value based on regression line.

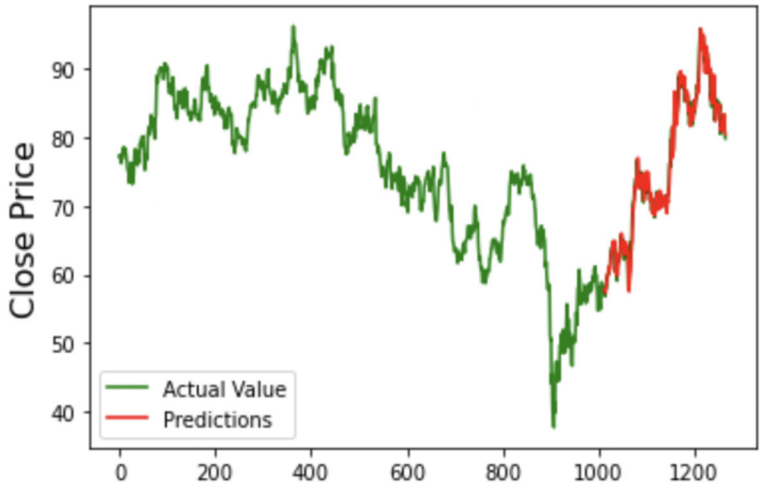

Fig. 3 Linear Regression prediction

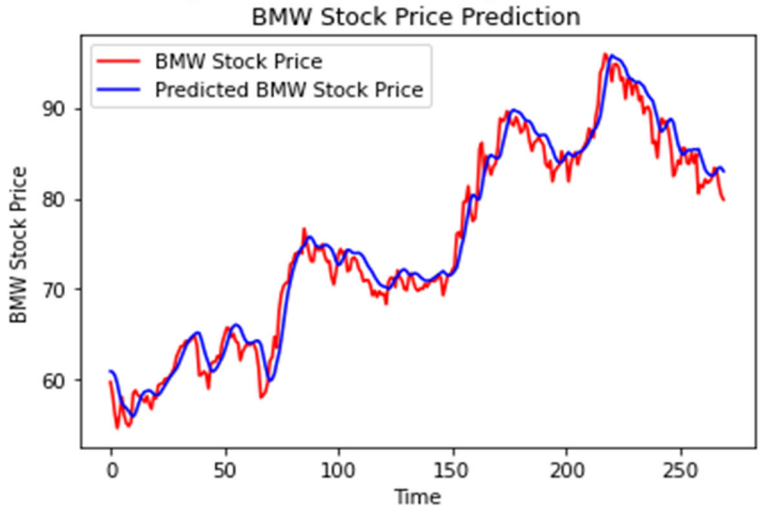

Fig. 4 LSTM prediction

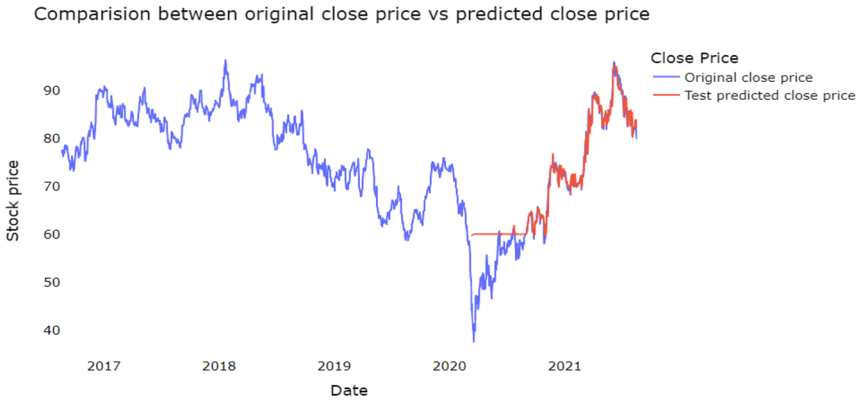

Fig. 5 Random Forest Regression prediction

Table 2. Comparison table.

	Multiple Linear Regression	LSTM	Random Forest
RMSE	1.238	1.99	5.02
MAE	0.939	1.54	2.87
R-squared	0.985	0.96	0.87

As depicted in Figs. 3-5, there are the three prediction model plot. By analyzing their statistics. A comparsion tables for the metics are listde in Table. 2. Accoridng to the results, the comparative analysis states that the Multiple Linear Regression present the best model. It indicates the highest R-squared and the lowest RMSE. Thus, it derives the conclusion that the multiple linear regression model will make the most accurate prediction on BMW stock price rather than LSTM and Random Forest.

4. Limitations & Future prospects

Predicting on stock market price is challenging, any detail that depends on the multiple independent variables is significant. Lacking other extra features will be the limitations and shortcomings of this article. This is because despite the basic open, high, low, close, adjusted close, and volumn that would show in the stock market, other features (e.g., oil price, opponent company stock situation, and current economical circumstances) might also strongly affect the price. In the future, the model could add more independent variables as the input features and improve the model to make a more precise prediction on BMW's stock price.

5. Summary

In conclusion, this paper investigates the BMW's stock price prediction based on machine learning. Specifically, regression modeling. According to the analysis, multiple linear regression model states the best accuracy and lowest error. Besides, random forest regression might have overfitting problem. This is because the training fit line presents a nearly 1.00 on R-squared; however, its testing dataset exhibits 0.87 R-squared. This might be attributed to overfitting on the training set. In the future, more effective factors can be taken into consideration, which will enhance the performances. Overall, these results offer a guideline for future stock market analysis based on stock price prediction.

References

[1] Shah D, Isah H, Zulkernine F. Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 2019, 7(2): 26.

[2] Obthong M, Tantisantiwong N, Jeamwatthanachai W, et al. A survey on machine learning for stock price prediction: algorithms and techniques. 2020.

[3] Hu Z, Zhao Y, Khushi M. A survey of forex and stock price prediction using deep learning. Applied System Innovation, 2021, 4(1): 9.

[4] Katherine Li. Predicting Stock Prices Using Machine Learning. Neptune.ai, 22 July 2022, Retrieved from: https://neptune.ai/blog/predicting-stock-prices-using-machine-learning.

[5] Bhattacharjee I, Bhattacharja P. Stock price prediction: a comparative study between traditional statistical approach and machine learning approach. 2019 4th international conference on electrical information and communication technology (EICT). IEEE, 2019: 1-6.

[6] Patel Janki, et al. Stock Price Prediction Using RNN and LSTM - JETIR. Nov. 2018, Retreived from: https://www.jetir.org/papers/JETIRK006164.pdf.

[7] Vijh M, et al. Stock closing price prediction using machine learning techniques. Procedia computer science, 2020, 167: 599-606.

[8] Chen W, Zhang H, Mehlawat M K, et al. Mean–variance portfolio optimization using machine learning-based stock price prediction[J]. Applied Soft Computing, 2021, 100: 106943.

[9] Retrieved from: https://www.autoevolution.com/news/bmw-and-lexus-pass-mercedes-benz-in-2015-us-sales-but-the-race-is-still-tight-103440.html

[10] Hiransha M, Gopalakrishnan E A, Menon V K, et al. NSE stock market prediction using deep-learning models. Procedia computer science, 2018, 132: 1351-1362.