# Stock price prediction using machine learning on least-squares linear regression basis



# Stock price prediction using machine learning on least-squares linear regression basis

**C. C. Emioma[1], S. O. Edeki [1*]**

[1]Department of Mathematics, Covenant University, Ota, Nigeria

Contact Emails: emiomacollins96@gmail.com, [*]soedeki@yahoo.com

**Abstract**. Predicting the future of a stock price is a difficult task due to the high level of randomness in the movement of prices. This research aims to use a machine-learning algorithm to estimate the closing stock price of a dataset to help aid in the prediction of stock prices leading to higher accuracy in prediction. The intention of the model is for it to be used as a day trading guide. The algorithm being used is called the least-squares linear regression model. It takes in a dependent variable, in this case, would be our closing price of the stock and an independent variable, which is the day each stock price was recorded.

Keywords: Stock price, machine learning, prediction.

## 1. Introduction

Machine learning refers to the study of algorithms and statistical models. It is a branch of artificial intelligence that gives computers the ability to learn from data and perform tasks without being explicitly given step by step instructions, only relying on the data it uses to train [1-3]. Machine learning is used in different applications, typically where it is hard to create a conventional algorithm for solving a problem. The stock market is a collection of buyers and sellers of stocks. A stock represents a fraction of ownership of a company by an individual or a group of people [4-6]. There are different ways to approach stock price prediction; some of them include technical Analysis, fundamental Analysis, time series analysis. With the progress in the development of technology, stock price prediction has moved into the technological space. Some of the best methods of prediction involve the use of recurrent neural networks (R.N.N.) and artificial neural networks (ANN), which also falls under the category of machine learning. Because machine learning falls under artificial intelligence, it allows the model to learn and improve from past experiences without having to be reprogrammed each time [7-9]. Some traditional methods for prediction using machine learning involved Backward propagation, otherwise known as back propagation errors. Nowadays, researchers are using different methods such as support vector machine (SVM) for predictions. Stock market movement for a short window seems to be a random process, but in the long run, it begins to

develop a linear curve. Because of this, investors always tend to buy stocks that they expect will increase in the near future. Uncertainty in forecasting stock values prohibits most individuals from investing in stocks. There is also a need to reliably forecast the stock price that can be applied in real-life situations. In a company's stock, the dataset includes information such as opening & closing price, highest price, lowest price, date, volume & adjusted closing price. In some algorithms, these variables are used for the prediction of the object variable, which is the adjusted closing price, but in this study using a linear regression algorithm, only the date variable will be used for predicting the stock price. The stock market contains information about all stocks available in the system. It is based upon different countries; for example, in Nigeria, we have the Nigerian stock exchange (NSE); in another country like the united states, they have multiple stock exchanges for particular states, i.e., NASDAQ is a New York stock exchange, B.X.Y. exchange is a Chicago stock exchange, etc. the stock exchange comprises of different sectors which include I.C.T., oil & gas, agriculture, energy, healthcare, real estate, financial services, etc. each of the sectors has various companies within them. Activities of a stock exchange include monitoring of opening & closing price, All-share index, equity cap, bond cap, volume, ETF cap for all it is listed stocks. Therefore there is a vast variation each day among every stock. Stock price prediction is at the forefront of attention for years because it has the potential to make large revenues. Predicting the stock market is a daunting task; it is a result of the near-random nature of the real-time series. The first two methods used to predict stock prices were the fundamental analysis method and the technical analysis methods, and now machine learning is also being used [10-12].

One of the ways to grow your income over time is to invest in the stock market, but people are hesitant to do so because of its seemingly random and volatile nature. The stock exchange may seem chaotic and volatile at first sight, but that's not the entire truth. If the stock market is closely observed, the pattern will slowly grow, and it will be clear that there are a variety of factors that lead to the company's stock price. Stock market prediction is defined as trying to determine the stock value and offer an accurate idea for the individuals to know the market. The ability to foresee the outcome of recent occurrences is a useful skill for investors. These occurrences can be economic growth, Interest rates, news, etc. these activities affect company earnings, and this, in turn, influences the market sentiment. It is beyond the reach of almost all investors to forecast such hyper-parameters correctly and consistently. All of these aspects make stock price prediction a difficult task. Individuals are most interested in buying stocks but are scared of the behavior of the stocks itself, thereby seeing it as a gamble. Once the right data is collected, it can then be used to train a machine and to generate a predictive result.

This research aims to predict the future price of a stock with a case study of the Bank of America stock. Series of researches on stock prices, predictions, machine learning, the text mining method, stock market behavior, and so on can be found in [13-15].

## 2 Presentation of Model

This section attempts to explain the necessary steps taken in the analysis and the tools used for the research. In linear regression, the relationship between the variables is modeled using linear predictor functions in which the model parameters are not known and are estimated from the dataset used [1-3].

This method of linear regression is called the least squares method. The formulas for the equations are as follows

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\left(y_i - \hat{y}_i\right)/\, y_i\right| \tag{1.4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{1.5}$$

The root mean squared error (RMSE) represents the standard deviation of the residuals, i.e., the difference between the model predictions and the actual values of the dataset. The mean absolute percentage error (MAPE) may exhibit some limitations if we have a data point value of zero because there is a division operation involved in the formula.

### *2.1 Data Collection*

The method of data collection is beneficial to any research work, and as such, there are methods and procedures for data collection. In this research, the data used was stock data of a financial institution. As it will not be possible to consider the whole financial stock, a selection was made, which runs for a period of 7 years. The dataset used was obtained online in yahoo finance from the Bank of America stock.

### *2.2. Missing values and used tools*

Missing values refers to the unavailability of some set of data point in the dataset and could affect the output of the research. In this research, there were no missing values
The programing tools employed for this study is a python programming language for training, testing, and validating the banking stock dataset used for the study.

## 3. Results and Interpretations

The result below shows the output of the price of the banking stock selected from the period of June 2012 to June 2019. The data was collected on a daily frequency, and classified into three parts: training dataset and test dataset and validating dataset on the proportion of 60%, 20%, and 20%, respectively.

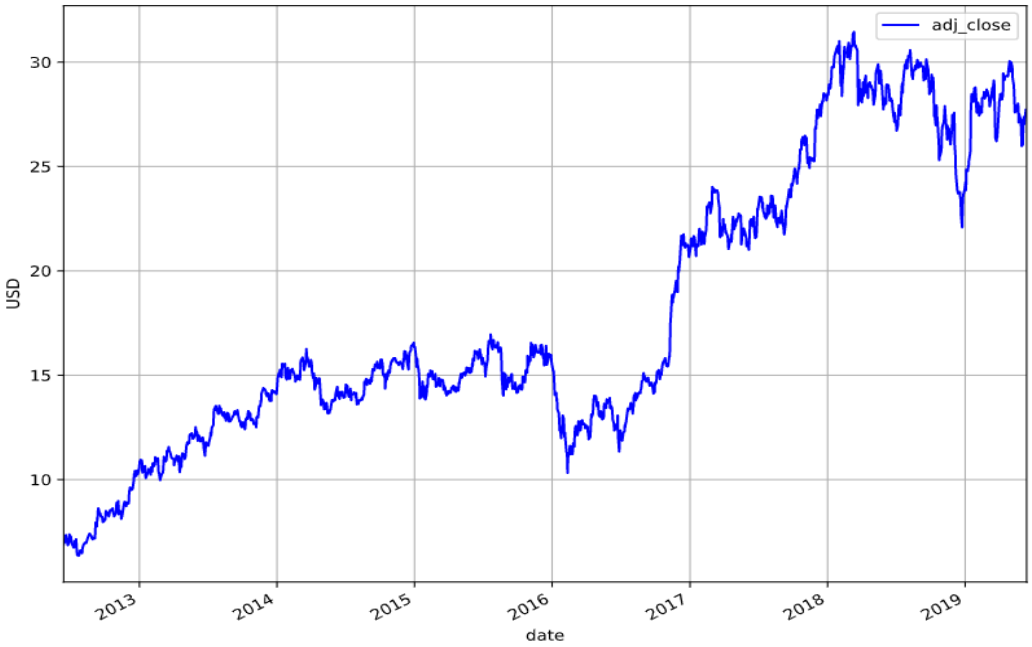



**Fig1**: Graph showing the plot of the stock price

The graph in Fig 1 shows the plot of the closing share price over the full-time period. The data set was split into three separate parts to enable the training of the model and getting accurate predictions. Also, it can be seen on the graph that there were seasonal variations in the closing price of stocks during the period. The stock price started with a bearish beginning during the year but kept following a bullish seasonal pattern.


Fig 2: Graph showing the splitting of the data set before training the model

In fig 2, before training can undergo, the dataset has to be split into three different parts, which are training, validation, and testing. The first section is used to train the model with past stock prices. The second section is used to evaluate the model and make changes to its parameters to increase its accuracy, and the final section is used to see the performance of the model. In the graph above, the training section is highlighted blue, the validation section is highlighted yellow, and the testing section is highlighted green.

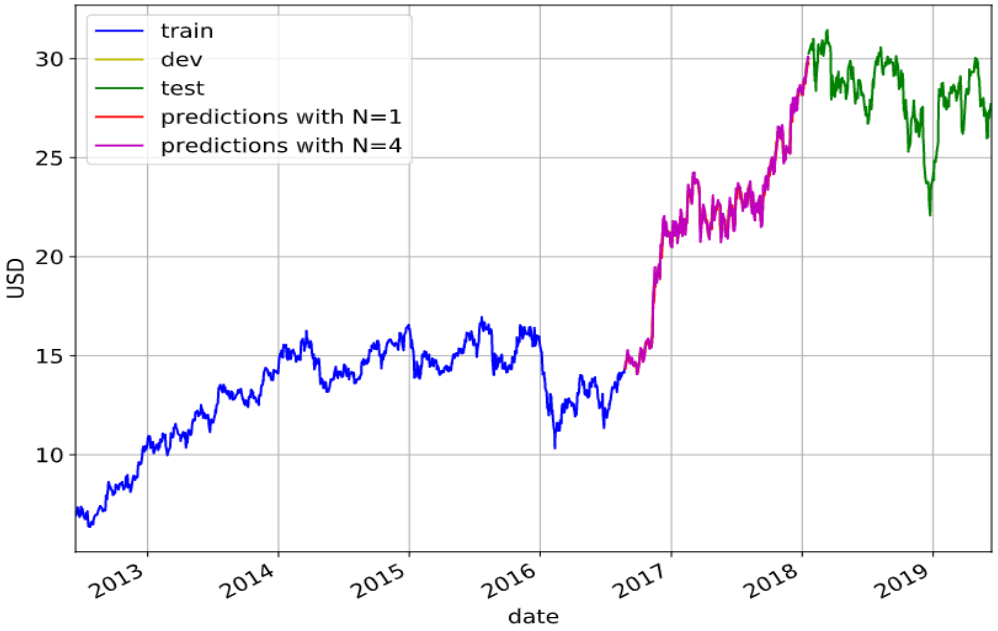
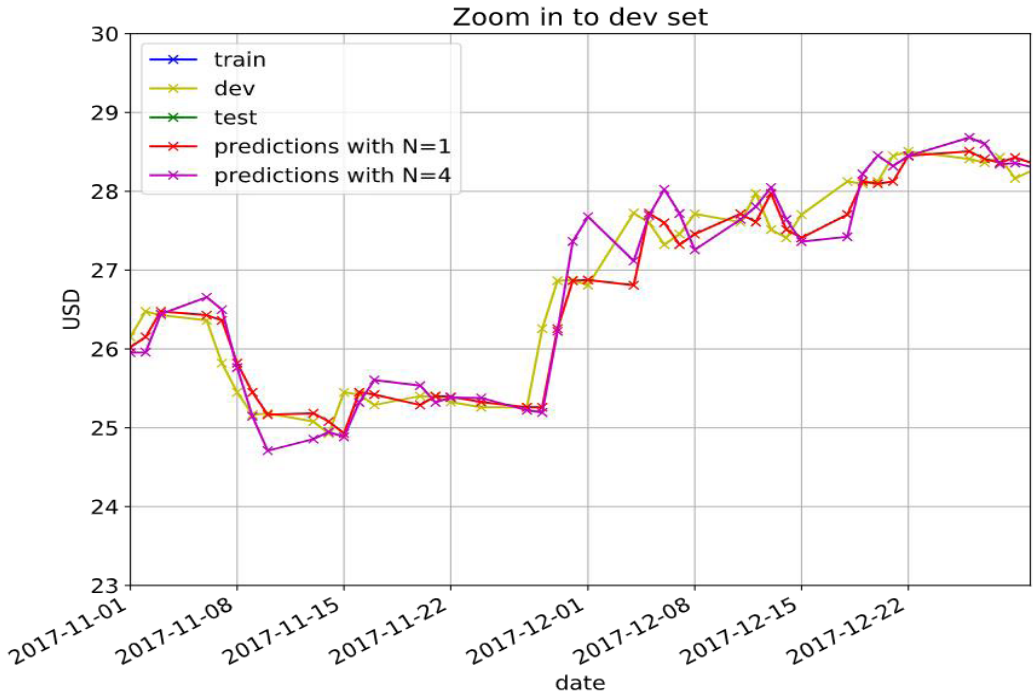



Fig 3: Graph showing the attempted prediction of the model on the validation data set
Fig 4: Post training of the model




After the training of the model Fig 4, it was used to predict the values of the validation section. The red and purple highlights show the predictions when the value of N is 1 and 4. N is the number of previous

days used to determine the present day's stock. N =4 achieved the lowest mean absolute percentage error, so it was used as the optimal value for prediction in the testing dataset.

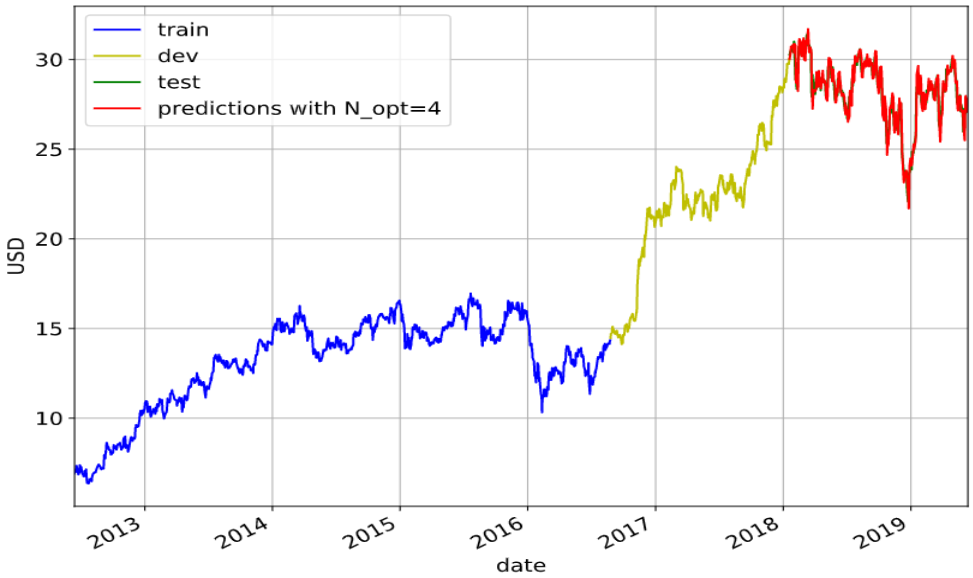



Fig 5: Trained model result 1

This graph shows the model prediction of the test dataset with the N optimal value set as 4. It achieved a mean absolute percentage error of 1.367%, and a root mean squared error of 0.512. the results were reasonably accurate. Therefore, this proves that the model can aid in the prediction of stock prices.

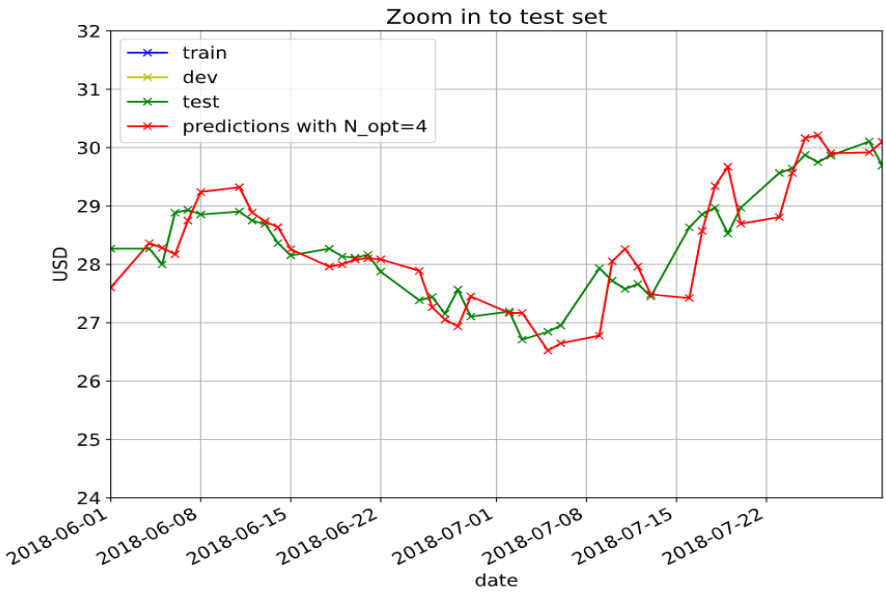



Fig 6: Trained model result 2

TABLE 4.1

| adj_close | Prediction |
|---|---|
| 30.26141 | 29.88934 |
| 30.47129 | 30.14693 |
| 30.45221 | 30.36635 |
| 30.6144 | 30.46175 |
| 30.6144 | 30.5 |
| 30.71934 | 30.6144 |
| 30.79566 | 30.66687 |
| 30.41405 | 30.7575 |
| 30.52853 | 30.60486 |
| 31.00554 | 30.47129 |
| 30.48084 | 30.76704 |
| 28.86855 | 30.74319 |
| 29.76532 | 29.67469 |
| 29.81302 | 29.31693 |
| 28.37246 | 29.78917 |
| 28.93533 | 29.09274 |
| 29.689 | 28.65389 |
| 29.74624 | 29.31216 |
| 30.52853 | 29.71762 |
| 30.72888 | 30.13739 |

## 4.Conclusion

In this work, the least-squares linear regression method was used for the prediction of the closing price of the bank of America stock dataset. The steps for the prediction include: Splitting the dataset into 3 (training, validation & testing datasets), Training the model with the training dataset, calculating the errors mean absolute percentage error (MAPE), root mean squared error (RMSE), Tweaking the parameters on the validation dataset to achieve the lowest errors, and Predicting the test dataset. The essence of the study was to create and test a machine learning model to aid in the analysis and prediction of a stock trend pattern. The model successfully predicted the results with a mean absolute percentage error of 1.367% and a root mean squared error of 0.512. The proposed model can be used by modifying only the training data

for any other stock market in other countries. With few changes, the model can be used for different purposes, such as analyzing and predicting the performance of students, predicting fuel consumption of a vehicle, and tracking the health of a patient, and many more. It is recommended to use the model for future predictions in the range of a day and update the dataset to make further predictions for the best accuracy.

**Acknowledgment**

All forms of support from Covenant University and constructive suggestions from the anonymous referee(s) are appreciated.

**References**

[1] Ahangar, R. G., Yahyazadehfar, M., Pournaghshband, H., (2010). The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange. International Journal of Computer Science and Information Security, 7(2), 38–46.
[2] Bikker JA, Spierdijk L, Hoevenaars RP, Van der Sluis PJ, (2008). Forecasting market impact costs and identifying expensive trades. Journal of Forecasting. 27(1). 21–39. doi: 10.1002/for.1052
[3] Dzikevičius, A., Šaranda, S. (2010). E.M.A. versus S.M.A. usage to forecast stock markets: The case of the S&P 500 and O.M.X. Baltic Benchmark. Business: Theory and Practice, 11(3), 248-255. https://doi.org/10.3846/btp.2010.27
[4] Edwards, R. D., I. Magee, J. (2007). Technical Analysis of Stock Trends, 9.2. Introduction to Stock Charts, (2009). Learn Stock Options Trading.com
[5] Janssen, C., Langager, C., & Murphy, C. (2012). Technical Analysis: Indicators and Oscillators.
[6] Naeini Pakdaman, Mahdi, Taremian, Hamidreza, B. Hashemi, Homa, (2010). Stock market value prediction using neural networks, 132 - 136. doi:10.1109/CISIM.2010.5643675.
[7] Adebiyi, A.A., Adewumi, A.O., Ayo, C.K., Stock price prediction using the ARIMA model, 2014 Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014 7046047, pp. 106-112.
[8] Adebiyi, A.A., Adewumi, A.O., Ayo, C.K., Comparison of ARIMA and artificial neural networks models for stock price prediction, 2014 Journal of Applied Mathematics, 2014,614342
[9] Kyoung-jae Kim (2003). Financial time series forecasting using support vector machines, neurocomputing, 55, 307-319. https://doi.org/10.1016/S0925-2312(03)00372-2.
[10] Mieko Tanaka-Yamawaki, Seiji Tokuoka, Keita Awaji (2009). Short-Term Price Prediction and the Selection of Indicators, *Progress of Theoretical Physics Supplement*, *179*, 17–25. doi:10.1143/PTPS.179.17
[11] Nison, S. (1991). Constructing the Candlesticks, In *Japanese Candlestick*, 21-26
[12] Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S., (2010). Text mining approaches for stock market prediction. In Computer and Automation Engineering. 4, 256-260.
[13] Olaniyi, S.A.S. & S., Adewole & Jimoh, R., (2011). Stock trend prediction using regression analysis-A data mining approach. ARPN Journal of Systems and Software. 1. 154-157.
[14] Rinehart, M. (2003). Overview of Regression Trend Channel (R.T.C.), 1-10
[15] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J., (1998). Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics,* 3, 2720-2725.