Price Prediction for Starbucks Corporation Based on Random Forest, Linear Regression, and Decision Tree

Xiaozhu Zhang[1, *]

[1]School of Business, George Mason University, Fairfax, United State

*Corresponding author: xzhang45@gmu.edu

Abstract. The outbreak of the virus COVID-19 in 2019 has dealt an unprecedented blow to the global economy. Contemporarily, the global economy is in the recovery stage after the end of the epidemic, and the economic recession is likely to be the future trend in terms of current status. In this paper, three state-of-art multivariate supervised learning regression scenarios (i.e., Random Forest, Linear Regression, and Decision Tree) are implemented to predict the stock price of Starbucks. Through comparison, it is concluded that Linear Regression and Random Forest are the most accurate models that are expected to be used in the future. Moreover, investing in Starbucks stock is not a wise choice at this time. It is hoped to reduce investor losses by choosing the optimal model for the most accurate stock predictions and minimizing losses in a recession. These results shed light on guiding further exploration of targeted analysis of the new economic environment.

Keywords: Stock Price; Starbucks; Decision Tree; Linear Regression; Random Forest.

1. Introduction

Predicting and analyzing the financial markets in context of different models are a crucial challenge even based on the most state-of-art techniques. This can be attributed to plenty of factors, for example, the stochastic volatility and a large number of possible effects that influence the financial market. Any stock market expert will find it very challenging to accurately predict market growth and collapse based on these considerations. Contemporarily, with the development of the machine leaning techniques, various of approaches have been proposed to predict the market with abundant factors and features [1]. One of the most effective weapons in any company's arsenal is price prediction through machine learning. Because automating price forecasting can help investors stay up-to-date with your industry and ultimately improve the efficiency of your sales process [2].

Starbucks, the largest coffee shop network in the world, operated 33,833 stores, where about half of them were located in the United States. In the U.S., more than 8,900 Starbucks stores are owned and controlled by the company, and the rest are licensed to operate [3]. Starbucks has been focusing on its digital apps after the pandemic. In the second quarter of 2022, it had 26.7 million active users. The pandemic and inflation have sent Starbucks stock down 35% over the past six months. While the index has been falling over the same period, investors are finding it difficult to regain trust due to rising commodity costs. Starbucks' U.S. same-store sales rose 12% in the second quarter, while overall sales rose 15% year over year. Profit margins have been shrinking recently due to inflation. In addition, their buyback program was halted. They have opened a new "pick up" store in Canada and adopting these advanced procedures and technologies will help them solve this problem. The National Labor Commission is asking Starbucks to reinstate seven employees who were fired for union activity. What modern investors care about is the social status a business takes [4].

Some people have studied the comparison of the ANN model and RF model [5] and the combination of the ANN model and DT model. There is no doubt that ANN models are the preferred choice of researchers, which gives RMSE (0.42), MAPE (0.77), and MBE (0.013). There is also a report comparing the LSTM model with the traditional MA model [6, 7]. The results indicate that LSTM is more predictable than traditional MA, but more methods need to be mined.

Therefore, this study wants to study separately and compare the DT model and RF model, at the same time, study a new model-Linear Regression. Now the global economy is in the recovery phase after the end of the epidemic, but an inverted yield curve appeared in the previous few months and heralded a recession. Prediction of stock prices is particularly important at this time. Therefore, the

best model is pursued for the stock price prediction of the same company through several models. In the three sets of models, the same set of data will be used to make predictions and use some indicators to test the accuracy of the predicted value. Finally, the optimal model and the future trading trend of Starbucks stock are determined by comparison. The rest part of the paper is organized as follows. The Sec. 2 will introduce the model used in this paper. The Sec. 3 will detail the steps to run each model. The Sec. 4 will point out the limitations. Eventually, a summary of the whole study and implications of the results will be presented in Sec. 5.

2. Data & Method

2.1 Data

The data used for this study is collected from finance.yahoo.com, which is stock information about Starbucks Corporation from 2017-07-19 to 2022-08-05. The objects of the data set include Date, Open, High, Low, Close, Adj Close, and Volume. The visualization of the price trends is given in Fig. 1. The utilized features (x) and prediction target (y) used in the three groups of models are previous close price and close price, respectively.

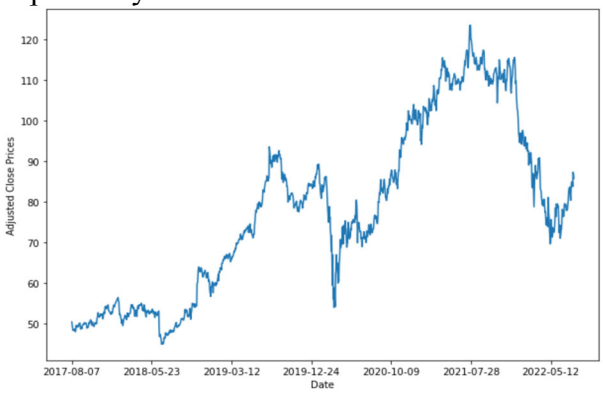

Fig. 1 Starbucks stock close price trend chart

2.2 Models

In this study, three models of machine learning will be adopted to predict the stock price of Starbucks Corporation, i.e. Linear Regression, Decision Tree, and Random Forest. As a matte fact, machine learning concepts are widely adopted in a broad range of cases, e.g., speech recognition, computer vision, and medicine, when creating traditional algorithms to perform the desired function is challenging or impractical. To be specific, in some kinds of analysis, schoalrs treat data and neural networks to a very similar way to real brain function of human beings [8].

Among various types of ML scenarios, the regression analysis in the sub-category of supervised learning is one of the most utilized apporach in exploring the relationship between different features and variables. The work of predicting the value of the dependent variable (y) based on the independent variables is done using linear regression (x). So with this regression method (output), x (input) and y (output) are linearly related. Hence, the term "linear regression" was born. The hypothetical function for linear regression is as follows:

$$y = \theta_1 + \theta_2 \cdot X \tag{1}$$

In this case (supervised learning), X is the input training data matrix (univariate - an input variable (parameter) label to the data). The best line fit by this model predicts the value of y for the given x during training. Find the best value of 1 and 2, where 1 is the intercept and 2 is the coefficient of x that allows the model to provide the best regression fit line [9].

Decision tree is a widely used category of supervised learning, where it splits and separates the data at differen nodes (also known as leaves) so that to realize various output [10]. As a updated version, the Random Forest approach can be regarded as the combination of the multiple decision tress. To be specific in othe words, it derives the final results based on baggering (e.g., average) the

results of different trees. Integrating meta-algorithms through bagging improves the accuracy of machine learning algorithms [11, 12]. A sketch of the algorithm of it is schematically presented in Fig. 2. The parameters and assessment methods for the three models are listed in Table. 1.

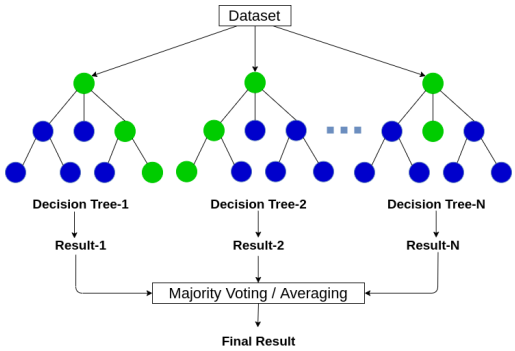

Fig. 2 A sketch of the random forest.

Table 1. Parameters and assessment method for each model.

Parameters	Assessment Method
• random_state=30 • min_samples_leaf=1 • min_samples_split=2 • max_depth=11 • n_estimaters=1000 • bootstrap=FALSE	• MEA • MSE • The R squared

2.3 Procedure

To get started, this study imported models that would be used in the following process: pandas, Linearregression, matplotlib.pyplot, statsmodles.api, pandas_ta, r2_score, mean_absolute_error, mean_squared_error, RandomForestRegressor, and DecisionTreeRegressor. Then this study used several functions to get metrics and figure, so that the next three models can calculate metrics and plotted predictions directly using the defined formulae: get_metrics(ols_pred), and get_plot(pred, t). Next, the data of historical stock price of Starbucks from 2017-07-19 to 2022-08-05 that downloaded from the finance.yahoo.com was loaded into memory as a panda data frame. The data had a new column called "EMA 10," which is the newly generated value. The following step was to obtain the EMA, and lagged values of volume, price, and EMA_10 (only use data before today for prediction). The last step of the preparation was test-train split. The train data was defined from 2017-07-19 to 2022-08-05, while the test data was defined One year before 2022-08-05. After the preparation, the study run linear regression model first. By estimating the model, it got the OLS regression Results. Then three metrics (R^2, MAE, and MSE) were calculated by the model using the formula defined above.

The study run decision tree model. First it estimated the model by using the parameters as mentioned in Table. 1. Then, it fitted with the train data. By using formulas defined above, the model calculated its metrics and plot the compare image between actual price and predictions. Finally, this study run random forest model with the same parameters with decision tree model and then used formula defined to calculated metrics and plot the compare graph. At the end of the study, predictions of three models for one year were put together in a prediction comparison graph while their metrics were put in a comparison table so that the difference among those three models could be seen clearly.

3. Results & Discussion

3.1 One-year predictions

The result of one-year prediction for Linear Regression model are depicted in Fig. 3. The predicted results are basically in line with the actual values in the one-year period. The result of one-year prediction for Decision Tree model are showed in Fig. 4. The fit of this model is not very good.

Although the evolution trend of the predicted value and the actual value is similar, the gap seems to be large, which is especially evident at the extreme point. The biggest point of error was in March 2022. The result of one-year prediction for Random Forest model are showed in Fig. 5. From the results it comes out that it is very similar in accuracy to Decision Tree model. The reason for this is the Random Forest model is based on the Decision Tree model, additionally, they use same parameters. The difference between the two can be seen more clearly in the comparison chart in the next section.

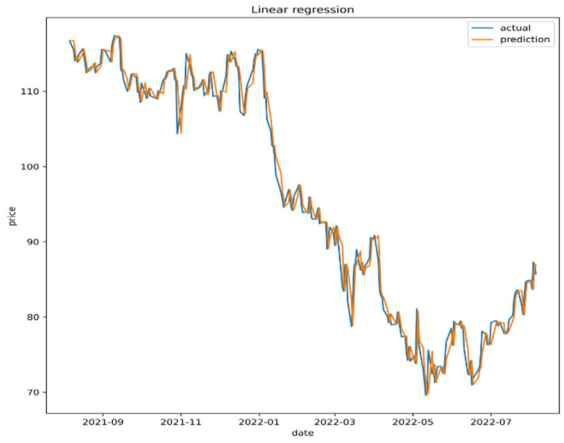

Fig. 3 Linear Regression model one -year prediction

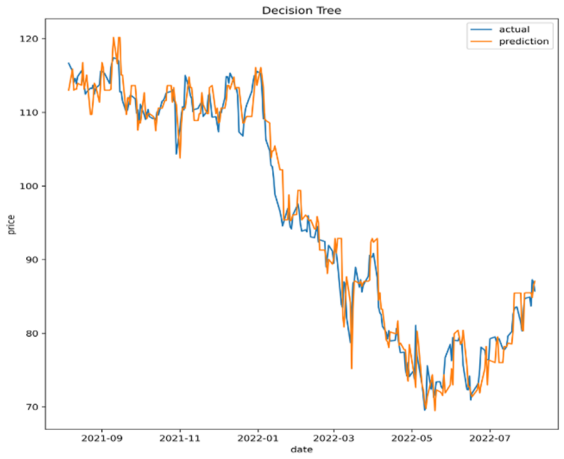

Fig. 4 Decision Tree model one-year prediction

Fig. 5 Random Forest model one-year prediction

3.2 Prediction Comparision

For a more celebratory comparison of the predictions of the three models, especially Decision Tree model and random Forest model, this paper put their predictions in one graph as shown in fig. 6. It can be seen from this graph that the model best fits the actual value is the Linear Regression model. There is a small gap between the Decision Tree model and Random Forest Model from January to March 2022, but after March 2022, the forecast results of the two models basically overlap. By comparison, the Random Forest model has slightly higher accuracy than the Decision Tree. However, the trends for those three models are the same with the actual price. The outbreak of the Covid-19 epidemic had caused Starbucks' stock price to plummet, and its stock price is improving at the end of the epidemic, but it is far from its previous level.

Fig. 6 Three models one-year predictions

3.3 . Metrics Comparision

To measure the accuracy of the model we apply three metrics: R suqared, MAE, and MSE, which is showed in Table 2. For multiple regression, it is sometimes called the coefficient of determination or multiple coefficients of determination. It is calculated as by $R-squared = Explained\ variation/Total\ variation$. MAE calculates the average residual value of the dataset. The calculation formula is $NAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_{pre}|$, where y_{pre} represents the predicted value of y. MAE is a negative score, the lower the better [13]. MSE quantitatively evaluates the variance of the residuals, which is calculated by $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_{pre})^2$. It is also a negatively oriented score, so the lower the better. According to the analysis (seen the results listed in Table. 2) the Linear Regression model has the largest R square, smallest MAE and MSE. It performs the best in this study, especially a lot ahead in MSE. According to the metric, the Random Forest model is more accurate than the Decision Tree model.

Table. 2 Three models compare metrics

	R squared	MAE	MSE
Linear Regression	0.986442	1.316964	3.273256
Decision Tree	0.973980	1.850428	6.281898
Random Forest	0.974927	1.803508	6.053244

4. Limitations and future prospects

The condition number for the linear regression model is huge, 6.52e+07. This might be a sign of significant multicollinearity or other numerical issues that would impair prediction accuracy. Then, the Decision Tree model and Random Forest are too similar with each other, there maybe other factors that out of the model will affect the results. With this in mind, this paper recommend to optimize the

Linear Regression model and choose different parameters between the Desicion Tree model and Random Forest model or choose another model like LSTM model for fruther study.

5. Summary

In conclusion, this paper investigates stock price prediction of Starbucks based on machine learning. According to the analysis, the Linear Regression model is the best choice for the future. In addition, the Decision Tree model and the Random Forest model have relatively large errors. For Starbuck's stock price analysis, the 2019 pandemic hit its stock price hard. Although it has improved towards the end of the epidemic in 2020, it has not returned to its previous level. Looking at the forecasts, all three models show that future stock prices will be highly volatile and have low prices. Investing in Starbucks Corporation stock now carries a large risk of loss, which is not an intelligent choice. Besides, the results of the three models are complementary, and the results of these three models are usually combined to get a more accurate prediction. So, it is suggested to predict the stock price of a company by using multiple models or a combination of models, which will greatly improve the accuracy and reduce the risk of investment. It is hoped that future research can improve the research method and the unified evaluation index, which will make stock forecasting more accurate. Overall, these results offer a guideline for stock price prediction with several models in machine leaning.

References

[1] Strader T J, Rozycki J J, Root T H, et al. Machine learning stock market prediction studies: Review and research directions. Journal of International Technology and Information Management, 2020, 28(4): 63-83.

[2] Price Prediction: How Machine Learning Can Help You Grow Your Sales. DLabs.AI, 8 June 2022, Retrieved from: https://dlabs.ai/blog/price-prediction-how-machine-learning-can-help-you-grow-your-sales/.

[3] Starbucks Corporation (SBUX) Company Profile & Facts. Yahoo! Finance, Yahoo!, 31 Aug. 2022, Retrieved from: https://finance.yahoo.com/quote/sbux/profile/.

[4] CrowdWisdom360. “Starbucks Stock Forecast: What Is Happening in Union Battle?” CrowdWisdom360, 21 July 2022, Retrieved from: https://crowdwisdom.live/us-stocks/starbucks-stock-forecast/.

[5] Vijh M, Chandola D, Tikkiwal V A, et al. Stock closing price prediction using machine learning techniques. Procedia computer science, 2020, 167: 599-606.

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

[7] Tsai C F, Wang S P. Stock price forecasting by hybrid machine learning techniques. /Proceedings of the international multiconference of engineers and computer scientists. 2009, 1(755): 60.

[8] Machine Learning: What It Is and Why It Matters. SAS, Retrieved from: https://www.sas.com/en_us/insights/analytics/machine-learning.html.

[9] ML: Linear Regression. GeeksforGeeks, 18 May 2022, Retrieved from: https://www.geeksforgeeks.org/ml-linear-regression/.

[10] Decision Trees for Classification: A Machine Learning Algorithm. Xoriant, Retrieved from: https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm.

[11] Mbaabu O. Introduction to random forest in machine learning. Berreskuratua-(e) tik, Retrieved from: https://www. section. io/engineering-education/introduction-to-random-forest-in-machine-learning, 2020..

[12] Zhao X, Wu Y, Lee D L, et al. iforest: Interpreting random forests via visual analytics. IEEE transactions on visualization and computer graphics, 2018, 25(1): 407-416.

[13] Editor, Minitab Blog. Regression Analysis: How Do I Interpret R-Squared and Assess the Goodness-of-Fit? Minitab Blog, Retrieved from: https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.