Stock Price Prediction Based on Simple Decision Tree Random Forest and XGBoost

Yifan Wu *

Statistics and Data Science, Carnegie Mellon University, Pittsburgh, US

*Corresponding author: wu3@andrew.cmu.edu

Abstract. Stock price prediction is the process of using a variety of methods and techniques to forecast the future performance of the stock. It is helpful for investors and traders to make decisions before and during trading. In this study, three stock prediction models depending on the decision tree algorithm are developed (i.e., Decision Tree, Random Forest and XGBoost) To be specific, their performances are compared to select the best prediction approach. Both accuracy and mean squared error (MSE) are utilized as metrics to evaluate the performances of models. By comparing the results and the metrics calculated according to the predictions, the random forest model is the best model with the highest accuracy and the smallest MSE. However, the XGBoost was the least accurate since this algorithm is too sensitive to the scale of the input variables. These results offer suggestions for choosing stock price prediction models and shed light on improving the models.

Keywords: Stock price prediction; decision tree; random forest; XGBoost.

1. Introduction

The stock market is the place where stock issuers are able to raise capital and the buyers can invest with their money. According to Mian, Muhammad, and Usman [1], the stock market influences the economy of a country by performing a critical function in the development of industry and commerce of this country. For the investors, the way they benefit is to buy when the stock price is low and then sell it when the price gets high. Therefore, the stock price and its trend are significant for the investors, and stock price prediction was born to forecast and utilize the trend. Most individual traders look at the indicators such as moving averages to predict stock prices. However, as Vijh et al. explained [2], stock price forecasting depends on a wide range of factors, making it a provocative task. Other scholars argued that it is extremely difficult to detect and analyze the impact of all factors on the stock price manually, people combine stock trading with computer algorithms and technologies such as machine learning (ML).

Random forest, XGBoost, and decision tree are three popular machine learning methods for stock price prediction. According to Myles et al. [4], the decision tree algorithm is divided and conquer for detecting attributes and generalizing regularities of data, which is useful and significant for modeling and forecasting. In the financial market, there are many stock price prediction models using the simplest decision tree. An example is a prediction made by Kamble. The model predicted not only the short-term stock trend but also the long-term trend with some oscillators and indicators (e.g., MACD, RSI, KDJ, and BB). The results showed that Kamble's models using the decision tree can produce better accuracy in short-term trading [5]. However, the accuracy of the model decreases when it comes to predicting stock trends in long term, suggesting that relying on these technical indicators is not sufficient.

As Loke stated, a random forest combines the use of a number of sets of decision trees [6]. Therefore, it is an ensemble classification algorithm. Khaidem, Saha, and Dey [7] predicted stock prices with the model using the random forest classifier. Since slight noise from the data may lead to a decision tree growing in a completely different way, the random forest algorithm was used to overcome this problem in their model. They trained multiple decision trees on many subsets, and it turns out that the model is indeed solid in predicting the stock trends, with a cost of a minor increase in bias. In terms of metrics, their model can achieve 85% to 95% long-term prediction accuracy for the datasets of stocks selected, namely Apple, Microsoft, and Samsung.

Similar to the random forest, XGBoost is also a tree-based classifier. It is short for eXtreme Gradient Boosting [8]. Besides, it is an open-source implementation of the gradient boosting algorithm based on the decision tree method [9]. According to Basad et al., both random forests and XGBoost are used to predict stock prices, and the models they developed demonstrated high accuracy in prediction [10]. Among the results, the random forest shows that the accuracy increases with the width of the window, while the stock price trends predicted using XGBoost resemble that of the random forest.

In the presented paper, to compare the accuracy of random forest, XGBoost, and decision tree in stock price prediction, models using the three methods listed above are created and improved on existing algorithms, and the results are shown. The three-stock data used and the details of the three models are described in the Data and Method section. In the Results and Discussion section, there are simple correlation analyses, a demonstration of the performances of different models, and a comparison. For the Limitation and Future outlooks section, the limitations of this paper and the future steps to be taken are elucidated. Finally, the Conclusion section includes the findings and a summary of the paper.

2. Data & Method

To evaluate the stock price prediction models, stock price data is needed. Therefore, stock prices of three companies from January 1, 2021, to July 1, 2022, are chosen, and they are Bank of America Corporation (BAC), Chevron Corporation (CVX), and The Home Depot, Inc (HD). The three companies come from three different industries. Bank of America Corporation is in the financial industry, while Chevron Corporation is an automotive company, which belongs to the energy industry. And The Home Depot, Inc. is a retailer of home decor and building materials. The stock prices are provided by Yahoo Finance and are downloaded using the yFinance package in Python. Fig. 1 shows the historical prices (adjusted close prices) of these three stocks for 18 months (from 2021-01 to 2022-07).

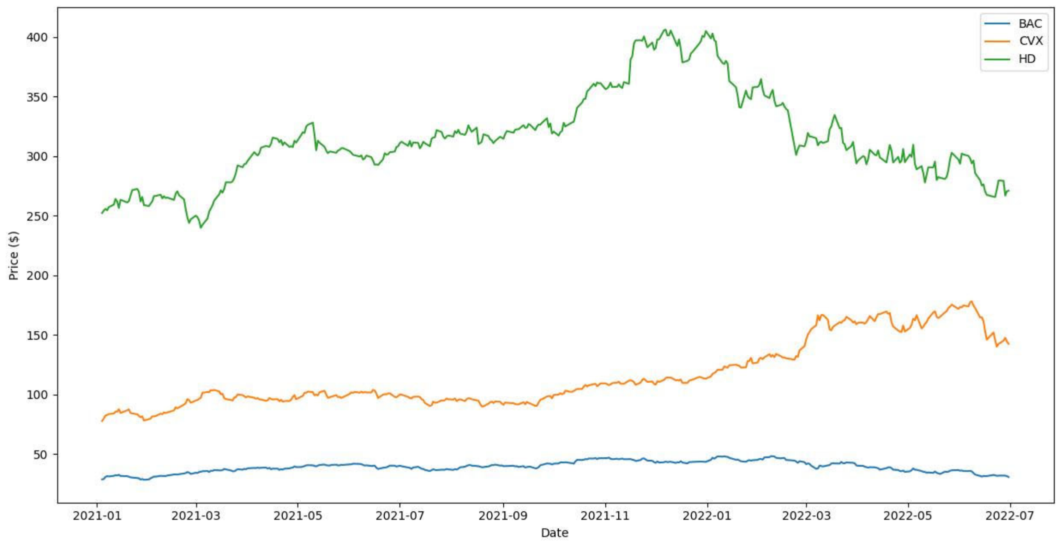

Fig. 1 History of stock price.

The variables used to predict stock prices are the open prices, the close prices, the low prices, the high prices, and the volume. The decision tree is a basic machine learning method, and in the presented paper, the stock price prediction model using the simple decision tree refers to using a single decision tree. It trains the model with the information of the stock including the historical price and the volume. After the assessment, the model is used to predict future stock prices. When training the model, the decision tree algorithm creates a tree structure where each internal node represents a feature-based decision, while each leaf node is a prediction. To create the tree structure, it divides the data into smaller subsets depending on the decision rules based on entropy, while the aim of the decision tree is to minimize the total entropy. In addition, as a way of using decision trees, a random forest is a combination of multiple decision trees rather than a single decision tree, which helps to improve the accuracy of the prediction.

XGBoost is also a machine learning method. Differing from the previous two algorithms, it is an implementation of gradient boosting. The XGBoost model uses the decision tree as its fundamental learner, and then optimizes the prediction with the following formula of the objective function:

$$Objective = \Sigma[L(y_{true}, y_{pred}) + \Omega(f)] \quad (1)$$

where L(y_true, y_pred) is the loss function that measures the error between y_true and y_pred, $\Omega(f)$ is the regularization term, which helps to prevent overfitting by penalizing complex models, and f is the predictive result of the model.

To evaluate the three models and to compare the predictions, the following metrics are used to estimate the models. Accuracy shows the percentage of correct predictions out of all predictions. The higher the accuracy is, the better the model is predicting. Mean squared error (MSE) measures the error, so a smaller MSE indicates a better stock price prediction of the model.

3. Results & Discussion

Before building the models, correlation matrix heat maps can be useful for screening valid predictors. Fig. 2, Fig. 3, and Fig. 4 depict the correlation matrix for the three stocks, respectively. The correlation matrix shows the correlation between different variables, and it is a measure of how closely the variables are related to each other.

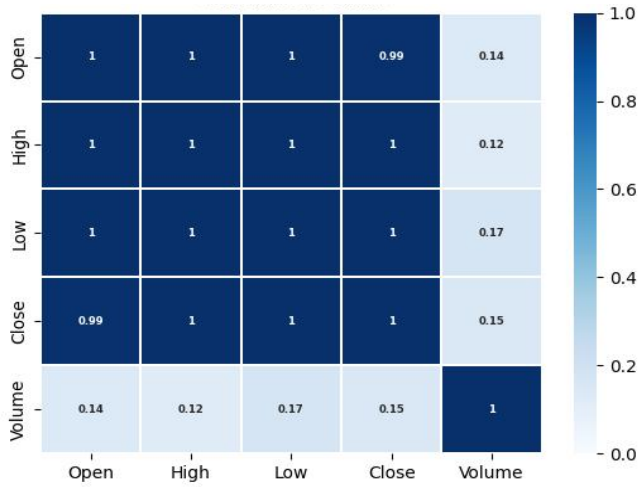

Fig. 2 Correlation Matrix for Bank of America (BAC).

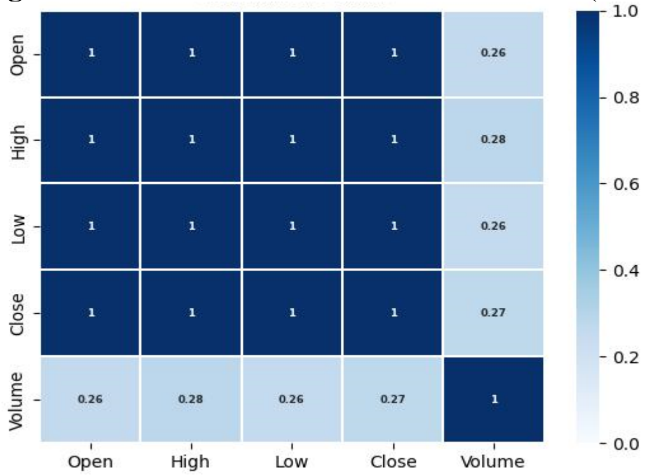

Fig. 3 Correlation Matrix for Chevron Corporation (CVX).

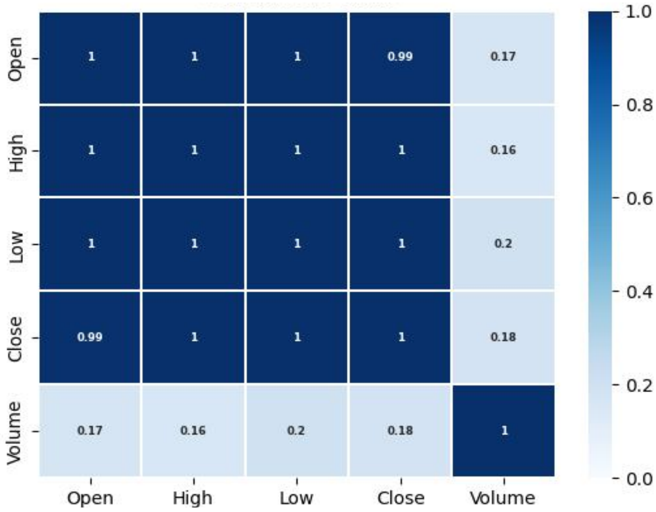

Fig. 4 Correlation Matrix for The Home Depot, Inc. (HD).

From the heatmaps of the correlation matrix given above, it can be easily seen that for these three companies, their open prices, close prices, low prices and high prices are a highly correlated set of data. Thus, these variables will be useful for constructing models to predict stock prices. For these three companies, the model using the decision tree algorithm applies the stock information, specifically the historical prices of the stock, as inputs and predicts the stock prices for the next ten days. The accuracy of the model for three stocks is shown in Table 1. Table 2 lists the MSE of this model. The stock price prediction model using random forests takes all kinds of stock prices and the volume of stocks as its input for training and predicting, and the predicted stock prices are compared with the actual adjusted close prices. The XGBoost model uses the same datasets as the previous two models. The corresponding accuracies for the random forest model and the XGBoost model are also given in Table 1, while Table 2 shows the mean squared errors.

Table 1. Accuracy of the predictions of the three models

Stock \ Model	Simple Decision Tree	XGBoost	Random Forest
Bank of America (BAC)	97.72%	95.19%	99.80%
Chevron Corporation (CVX)	98.15%	94.79%	99.74%
The Home Depot, Inc. (HD)	97.74%	94.11%	99.83%

Table 2. Mean Squared Error (MSE) of the predictions of the three models

Stock \ Model	Simple Decision Tree	XGBoost	Random Forest
Bank of America (BAC)	3.12	7.70	0.01
Chevron Corporation (CVX)	43.99	147.32	0.21
The Home Depot, Inc. (HD)	165.56	514.79	0.80

Seen from Table 1, the prediction accuracy of all three models is greater than 90%. Since accuracy is positively correlated with how well the models forecast, it means that these three stock price prediction models are effective for the selected stocks. The simple decision tree model is supposed to be the least accurate model, but the XGBoost model is the most inaccurate in its predictions. Furthermore, the prediction model using random forests has the highest accuracy, with all three stocks having a prediction accuracy of over 99%. Table 2 indicates the mean squared error (MSE) of the three models, and the model using random forests has the smallest MSE, which is 0.01, 0.21, and

0.80 for the three stocks, respectively. However, the simple decision tree model and the XGBoost model generate their large MSE, especially the XGBoost model. With the data in Table 2 and the observation of Fig. 1, all models have the property that their MSE value increases as the stock prices increase. The stock price of BAC is below $50, while the stock prices of CVX are slightly higher than it, in the range of $50 to $200. HD, on the other hand, has a stock price of more than $250. With the same trend as the stock prices, the MSE from low to high is BAC, CVX, and HD. In spite of this attribute being the same for all models, the models using XGBoost and the decision tree appear to be overly influenced by the stock prices. It suggests that their predictions are not as secure as the random forest model. Both metrics show that the model using random forests is the best model for stock price prediction, whilst the XGBoost model is the worst for the given data.

Stock price prediction can be used by investors and traders to make informed decisions. The introduction and measurement of three prediction models in the presented paper provide a basic evaluation of machine learning-based stock price prediction. The findings of the metrics show that the random forest model is the best choice among the three models for stock price prediction. However, in contrast, the results in the presented paper also expose the shortcomings of the models for predicting stock prices. It can be concluded from the results that none of the three models can achieve fully accurate stock price prediction, which suggests that in real life, investors may face losses if they apply any of these models to make predictions and trade depending on the prediction outcomes. Whereas, for the researchers investigating these algorithms and the prediction models, the presented paper can serve as a basis for them to explore and improve on the three decision tree-based models and to modify them or innovate new stock price prediction models according to the aforementioned problems.

4. Limitations & Prospects

According to the analysis, the XGBoost model gives a high prediction accuracy for the three stocks, but the MSE of this model is influenced by the stock prices used as the input values. If the stock price is high, the value of MSE is greater. This implies that the predictions of the model still have a significant bias. The reason can be attribute to the sensitive characteristics of for XGBoost algorithm to the scale of the input features. In addition, even though the purpose of the presented paper is to compare the stock prediction models based on the decision tree, it also generates limitations since the idea underlying all three models is the same machine learning approach (i.e., decision trees). Moreover, a concern of all three models is that they may be prone to overfitting, as this is a particular problem of the decision tree method. Besides, the data sample used to evaluate the accuracy and error of the models in the presented paper is small, with only three stocks and 18 months of data of each stock.

Therefore, in accordance with the constraints of the presented paper as described above, embellishment and expansion can be done for each limitation in the future. First, the size of the data sample can be increased, and more stock data and variables may be used to train the models and make predictions. To reduce the sensitivity of the XGBoost model to the magnitude of the inputs, it is a good idea to pre-process the data before training when the model is built in the future. The scale of the features can be tuned to zero mean and unit variance, which is the standard normal distribution. In the meantime, the problem of overfitting can be mitigated by using cross-validation or other techniques to select the appropriate tree depth. To make a comparison of the performances of more methods for predicting stock prices, it is feasible to build additional models that are not based on the decision tree, such as the neural network-based machine learning methods like RNN and LSTM.

5. Conclusion

In summary, this study investigates the stock price predictions based on three tree-based machine learning methods, i.e., the simple decision tree, random forest, and XGBoost. For the stock data used in training and predicting, the simple decision tree should have the lowest accuracy because the model only uses the basic decision tree idea. However, the XGBoost model turns out to be the one with the minimum accuracy, since it is affected by the scale of the input values. The stock price prediction using random forest runs as expected. It is a set of multiple decision trees and imposes more restrictions on the prediction rules, so the model predicts most accurately. Only tree-based models are built in the presented paper and the limitation of machine learning methods drives the models to overfit. For the future steps, the models will be refined to be more accurate. Meanwhile, stock price prediction using neural network algorithms can be investigated. This study compares existing models for predicting stock prices based on the decision tree, allowing one to choose the most profitable and least loss-prone method of stock price prediction, on which improvements and innovations can be made in the future.

References

[1] Mian S N, Muhammad M N, Usman J G. Relationship between economic growth and stock market development. African journal of business management, 2010, 4.16: 3473-3479.

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

[3] Obthong M, Tantisantiwong N, Jeamwatthanachai W, et al. A survey on machine learning for stock price prediction: algorithms and techniques. 2020, 1: 63-71.

[4] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6): 275-285.

[5] Kamble R A. Short and long term stock trend prediction using decision tree. 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2017.

[6] Loke K S. Impact of financial ratios and technical analysis on stock price prediction using random forests. 2017 International Conference on Computer and Drone Applications (IConDA). IEEE, 2017.

[7] Khaidem L, Saha S, Dey S R. Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003, 2016.

[8] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 2015, 1(4): 1-4.

[9] AWS. How XGBoost Works. Accessed 17 December, 2022. Retrieved from: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html.

[10] Basak S, Kar S, Saha S, et al. Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 2019, 47: 552-567.