Comparison Analysis of Stock Price Prediction Based on Different Machine Learning Methods

Zhiyuan Jiang[1](✉), Jiachen Liu[2](✉), and Lixuan Yang[3](✉)

[1] Statistics and Operations Research Department, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
Jamesjiang2323@icloud.com

[2] Computer Science and Mathematics, Wake Forest University, Winston-Salem, North Carolina, USA
15810144551@163.com

[3] Asian Institute of Digital Finance, National University of Singapore, Singapore, Singapore
e0809369@u.nus.edu

Abstract. This paper aims to compare the stock prices trend for industries affected by the pandemic in the post-Covid era. Specifically, the delivery and cardboard box industry are chosen as examples to project future growth in four different Machine Learning Methods. Furthermore, an optimized asset portfolio is constructed based on the asset Efficient Frontier Minimum Volatility Asset in order to provide a more precise projection of the stock prices. After the projection comparison, the Linear Regression Model fails to exhibit a logical trend. In contrast, the remaining three methods, Decision Tree, Random Forest, and Gradient Boosting Models, correspondingly, all show similar results and reasonably project the future growth.

Keywords: Stock Price Prediction · Machine Learning · Asset Portfolio

1 Introduction

The trends in the stock market sometimes get confusing. People, even scholars, often find it difficult to figure out a regular pattern of it [1]. There are too many factors involved in deciding a stock price, some of which are hard to be quantified, increasing the difficulty of modeling and predicting [2]. However, people still put much effort into this because a good prediction on stock prices can provide an instructive guidance for investors to make decisions, which is notably meaningful [3]. Moreover, under the influence of the Covid-19 Epidemic, people's living habits have changed in a far-reaching way. People reduce outdoor activities and spend more time at home [4]. Lots of on-site activities have followed the trend and become online, which made some industries profit from the epidemic [5]. For example, it's obvious to see that the delivery industry benefits a lot from Covid-19 since online shopping activities often involve the use of delivery services [6]. In this case, other industries such as cardboard boxes manufacture will also

Z. Jiang, J. Liu and L. Yang—These authors contributed equally.

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 59–67, 2023.
https://doi.org/10.2991/978-94-6463-198-2_7

profit from Covid-19 indirectly. This is because every time people deliver something, the first step they do is to pack all the stuff up, which will mostly use cardboard boxes [7]. Therefore, we made a conjecture that the cardboard boxes industry also benefits from the pandemic. This paper makes a stock price prediction on companies in these two industries with different machine learning methods.

Stock Price Predicting methods had been discovered and developed for several decades, and the most recent methods require techniques including Machine Learning. Previous studies [1, 3] have shown that there are different machine learning methods to predict the next-day trends of an asset portfolio, which include logistic regression (LR), artificial neural networks (ANN), support vector machine (SVM), random forest (RF), and so on [2]. And the study in this paper explored these methods and came up with the conclusion that SVM produced the best prediction performance with data collected from Kuala Lumpur Stock Exchange (KLSE). Besides, another study done by Yaohu Lin *et al.* combined the Chinese traditional eight trigrams with the latest machine learning methods to enrich and create a better model of stock prediction [8]. Bin Weng *et al.* proposed a new "financial expert system that can be used to predict 1-day ahead stock price" [3]. However, their works above are not completely applicable to our topic in this paper. Specifically, data is collected from different Stock Exchanges, which contain different companies from different regions. For example, Kuala Lumpur Stock Exchange (KLSE) is one of the largest stock markets in Southeast Asia, while Shenzhen Stock Exchange contains companies from mainland China [9]. Furthermore, companies from different regions have different characteristics and are influenced by different factors [10]. Therefore, the methods applied in other studies may not apply to our topic. Besides, most scholars focus on predicting the trends for a whole stock exchange instead of any one specific asset.

Unlike the previous studies shown in the above statement, it can be already deduced from Yahoo Finance that the influence is in accordance with our expectations due to the epidemic, and all the stock prices of FedEx, UPS, International Co., WestRock Co., and Packaging Corporation of America have increased from 7.29 2020 to 7.28 2021. Also, considering the cardboard box industry is a sector on which researchers rarely focus, our research provides fresh and newly available information to the broader public. Superior to other prediction models, the machine learning method is widely utilized, especially in manufacturing, to increase operational efficiency and lower costs. As a result, our research selects a machine learning prediction model to make our model most relevant to the industry. Besides, by mixing the cardboard box industry and delivery together, we create such multi-variety data that machine learning can handle and identify continuous trends and patterns than other prediction models.

Statements above come up with this new problem which is waiting to be solved: under the post-pandemic era, whether these companies can successfully sustain their momentum even after the pandemic passed or just lost the benefits and slump. Even though numerous scholars have done similar research on the topic of stock price predictions, where these researches have imposed a great significance on investors by providing them useful instructions and farsighted suggestions, the contribution of the research about this problem is still significant, as the industries that the research will be focusing on have never been reached before using Machine Learning methods.

2 Methodology

In this research paper, machine learning methods including Linear Regression, Decision Tree, Random Forest and Gradient Boosting are applied to predict the future stock price. 5 companies in total are selected where 3 of them are chosen from Cardboard Boxes Manufacturer and 2 of them came from Delivery Industry.

2.1 Data Preparation

As mentioned above, 5 stocks are selected, including International Paper Company, West Rock Co., Package Corporation of America, FedEx Corporation and United Parcel Service. These companies are all famous and in leading positions in their industries [11]. According to Yahoo Finance, the percentages of increase in stock price for some of them even exceeded 50%: UPS (55.27%), International Paper Company (59.18%), WestRock Co (64.44%).

All of the data are sourced from Yahoo Finance website and the timeline is set between 1st of January 2019 to 30th of June 2021 which covered the Covid-19 period from the beginning to the time where this research was based. Also, 5 sets of 525 data are enough for the Machine Learning process when allocating them into training sets and testing sets.

2.2 Asset Portfolio

Monte Carlo Simulation is applied for the determination of asset portfolio, whereby this approach all possible situations of asset portfolio for 5 stocks can be considered and conducted in one single graph.

$$\mathrm{E(X)} \approx \frac{1}{\mathrm{N}} \sum\nolimits_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{X_n} \tag{1}$$

where E(x) is the average of the random variable x, N is the number of trials in Monte Carlo Simulation.

The Efficient Frontier method is then used for further asset allocation. After simulation, the Maximum Sharpe Ratio Portfolio or the Minimum Volatility Portfolio is used for final asset allocation.

Maximum Sharpe Ratio Portfolio indicates that the investor is not satiable, which means either among the same risk, portfolio with higher returns will be chosen, or among the same amount of return, portfolio with lower risk will be obtained.

Minimum Volatility Portfolio indicates the measurement of the minimized price movement which leads to a greater chance of slow but steady returns over time.

The result will be determined by the results of Maximum Sharpe Ratio Portfolio and Minimum Volatility Portfolio based which will be analyzed further based on the objectives. It is also worth mentioning that the stability of the portfolio should be considered to avoid serious problems such as overweighting one stock in the portfolio.

2.3 Pre-processing and Model Fitting

Before the data is input for final prediction under Machine Learning Methods, the data should be pre-processed in order to separate into training set and testing set. Then 4 Machine Learning Methods are applied:

1) Linear Regression

This analysis typically produces results that are highly correlated to the variables. Ordinarily, the forecasted values are a linear combination of various inputs. For the prediction in this research, the output stock price is forecasted by various inputs based on a long period of time series [12].

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{2}$$

where y is the forecasted linear combination input of x, beta indicates the error between the predicted and actual value of y, sigma is the least square error.

2) Decision Tree

This analysis allows prediction to data by following the decisions in the tree from the root down to a leaf node. A tree consists of branching conditions where the value of a predictor is compared to a trained weight. The number of branches and the values of weights are determined in the training process. Moreover, modification may be used during the training process in order to simplify the model [13]. An important method Chi-square is introduced which is a classification method of the decision tree:

$$y^2 = \sum \frac{(O - E)^2}{E} \tag{3}$$

where O is the observed score, and E is the expected score

Also, it is worth mentioning the Information Gain which is used to split the data with entropy:

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X) \tag{4}$$

where T is the target variable, X is the variable of split data, Entropy (T, X) is the entropy calculated after the data is split of X.

3) Random Forest

This analysis is always considered as a development method to the Decision Tree method, where the basic logic is the same. However, rather than having one data in each ‘leaf node’, Random Forest approach engages with selecting multiple subsets of data randomly and ‘bagging’ them to build a tree. For Binary tree [14]:

$$ni_j = w_j C_j - w_{left(i)} C_{left(j)} - w_{right(i)} C_{right(j)} \tag{5}$$

where ni_j is the importance of node j; w_j is weighted number of samples reaching node j; C_j is the impurity value of node j; $C_{left(j)}$ is child node from left split on node j; $C_{right(j)}$ is child node from right split on node j.

Table 1. Stock prices data

	Adj Closes	Volume	HL_PCT	PCT_change	Volatility	Moving_Average
0	87.42	2558083.77	2.00	−0.34	0.43	87.42
1	88.67	3303321.09	3.30	1.79	0.43	88.67
2	91.79	3407109.63	3.54	2.81	0.43	89.29
3	92.43	1854002.25	1.69	−0.05	0.43	90.96
4	92.80	1797058.51	1.63	0.26	0.43	92.34

4) Gradient Boosting

This analysis is based on the Regression Tree algorithm, where unlike Decision Tree, Regression Tree applies mean residuals at each terminal node of the tree to determine the next node of the tree. Gradient boosting allows the average gradient component would be computed and helps in predicting the optimal gradient for the additive model, therefore accurate the prediction even further [14].

$$h_m(x) = \sum\nolimits_{j=1}^{J_m} b_{jm} R_{jm}(x) \tag{6}$$

where $\mathrm{h_m}$(x) is the output, Jm is the number of ‘leaves’, $\mathrm{R_{jm}}$ is the region and $\mathrm{b_{jm}}$ is the value predicted in that region.

3 Results and Discussion

We construct an Efficient Frontier for our five assets in order to optimize the portfolio with the lowest volatility in terms of future prices estimation. The Efficient Frontier aims to allocate the appropriate amount of weights for each asset to make the optimized portfolio.

3.1 Efficient Frontier

In order to create an optimized portfolio, we create an efficient frontier that illustrates the Sharpe Ratio as well as the Volatility Ratio distribution given the same amount of risk. At each vertical volatility level, the portfolio shares the same amount of risk, but the return varies.

According to Fig. 1, the maximum Sharpe Ratio has the highest return 202.96%. However, even though the maximum Sharpe Ratio Portfolio obtains higher returns, the weights of “UPS” in the portfolio is as high as 86.64%, meaning that “UPS” itself almost takes up the entire portfolio. This concept does not comply with the idea of portfolio diversification. Plus, the main objective for us is not to determine the highest return for a given portfolio, but to predict the future prices. As a result, instead of choosing the maximum Sharpe Ratio portfolio, we decide to select the portfolio with the minimum volatility. This selection makes the most logical sense in terms of its stability rather than maximum return.

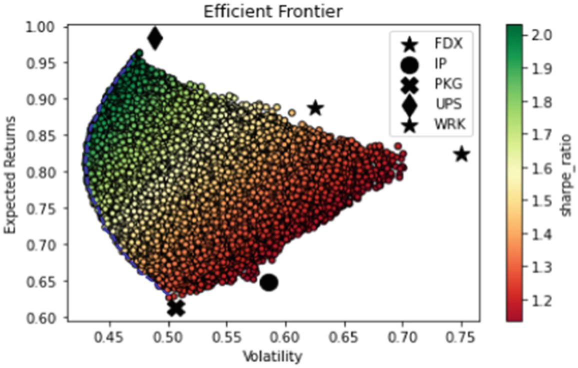

Fig. 1. The Efficient Frontier

3.2 Minimum Volatility Weights

The weights, correspondingly, are applied to the data in order shown in Table 1.

The weights, correspondingly, are applied to the data in order to form the table shown above. "Adj close" stands for the closing prices of the portfolio. "HL_PCT" indicates the percentage change of the highest and the lowest prices on a certain trading day. Volatility is the degree of variation of the portfolio's trading prices, and the moving average shows the trend of the prices in a given time period. With the help of these measurements, we are able to create a prediction model using different methods of machine learning, which are linear regression, decision tree, random forest, and gradient boosting.

3.3 Minimum Volatility Weights

Based on the result, we drew the prediction of the portfolio asset with 4 models run.

Figure 2 shows a general prediction from the testing set of the 4 machine learning methods. The main objective of this graph is to demonstrate the suitability of each

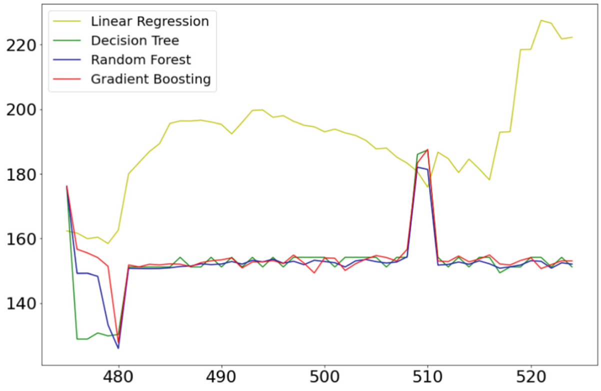

Fig. 2. Prediction of the Testing Set.

machine learning method. The less variant the line of the method, the more appropriate the method it will be. As a result, the three prediction models apart from Linear Regression model tends to become a relatively stable horizontal line, therefore, the other three machine learning methods may be more appropriate.

Figure 3 shows the testing set result from 4 machine learning methods for the future predicted period. The three other methods apart from Linear Regression show a quite steady constant return, which could determine that the other three methods are more appropriate for this approach.

Figure 4 shows the training set results from the 4 machine learning methods. The machine learning methods apart from Linear Regression show a much more accurate trend following the actual stock price with a lag. Therefore, the other 3 machine learning methods are more appropriate in this project.

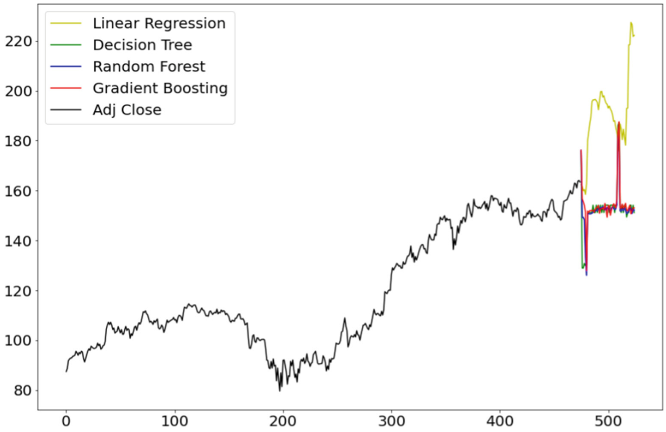

Fig. 3. Prediction with Machine Learning Methods

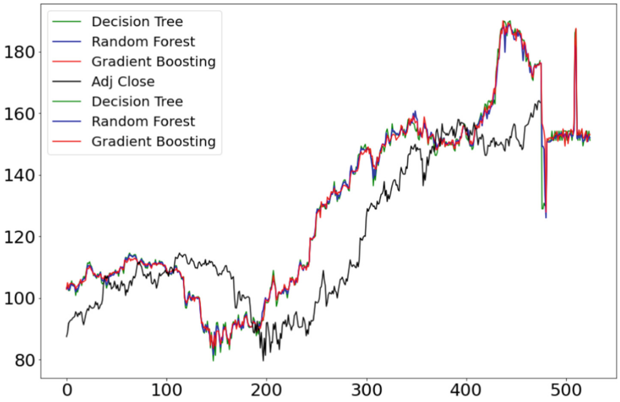

Fig. 4. Prediction of the Testing Set.

4 Conclusion

With the help of machine learning methods, 4 different results were obtained. Some of them show an overall optimistic trend whereas some show the portfolio assets will suffer from a drop at first but return to normal in the long run. In particular, the Decision Tree Method, the Gradient Boosting Method, and the Random Forest Method all show the trend of having a lag in projecting the stock price, whereas the Linear Regression Method fails to project a reasonable trend in terms of its unrealistic rising stock price. The trends based on the three qualified machine learning methods are similar. The stock price of the delivery industry and the cardboard box industry will experience a drastic fall after the recovery of the pandemic but will eventually bounce back to the level that is significantly higher than the average prices before the Covid-19. The fluctuations, meanwhile, become relatively stable after the recovery, showing no sign of a potential burst of stock prices or a plunge. The only exception of a drastic change of prices happens at the 500 level. Within a short amount of time, the stock prices increase about 30 percent and then drop back to their original level. Our assumption is that that another serious incident may occur in the close future, which will largely increase the demand for delivery and CBBs, but this event will then quickly be resolved as the demand seems to return to normal. Other than this distinctive change of the projection, the three optimal machine learning methods almost present the same trend for the future stock prices, becoming strong evidence regarding the accuracy of our prediction models.

References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. Shamim, Khalid, et al. “COVID-19 Health Safety Practices: Influence on Grocery Shopping Behavior.” Journal of Public Affairs, 2021, pp. e2624–e2624, https://doi.org/10.1002/pa.2624.
5. Lawton, Thomas C., et al. “The Implications of COVID-19 for Nonmarket Strategy Research.” Journal of Management Studies, vol. 57, no. 8, Wiley Subscription Services, Inc, 2020, pp. 1732–36, https://doi.org/10.1111/joms.12627.
6. Anupam Sharma, and Deepika Jhamb. “CHANGING CONSUMER BEHAVIOURS TOWARDS ONLINE SHOPPING - AN IMPACT OF COVID 19.” Academy of Marketing Studies Journal, vol. 24, no. 3, Jordan Whitney Enterprises, Inc, 2020, pp. 1–10.
7. Eric Roper. “Consumers Wrestle with Deluge of Cardboard Boxes from Delivery Services, Online Shopping.” Knight-Ridder/Tribune Business News, Tribune Content Agency LLC, 2017.
8. K. Elissa, “Title of paper if known,” unpublished.
9. Jiang, Zhuhua, et al. “The Effect of Air Quality and Weather on the Chinese Stock: Evidence from Shenzhen Stock Exchange.” Sustainability (Basel, Switzerland), vol. 13, no. 5, MDPI AG, 2021, p. 2931–, https://doi.org/10.3390/su13052931.

10. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
11. Rashid Mehmood Khan, et al. “Factors Influencing Stock Returns in Listed Firms of Karachi Stock Exchange.” Paradigms (Lahore, Pakistan), vol. 11, no. 2, University of Central Punjab, 2017, pp. 248–51, https://doi.org/10.24312/paradigms110219.
12. Belsey, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons, Inc.
13. Almuallim H., An Efficient Algorithm for Optimal Pruning of Decision Trees. Artificial Intelligence 83(2): 347–362, 1996.
14. Breiman, L. (2001). Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324