Stock Price Prediction of the TMT Industry Based on Machine-Learning

Tianyu Hu *

Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China

* Corresponding Author Email: tianyuhu@mail.ustc.edu.cn

Abstract. Predicting the TMT stock market enjoys great significance in the financial field and it has been widely researched due to its sophisticated nature and huge potential. Therefore, two machine learning models are trained and tested in this article, i.e., the multi-linear regression model and the random forest model with price information from three stocks, i.e., AAPL, NFLX, and TSLA. Every model used 80% of stock data to train and the rest 20% to test. To evaluate how accurately the models fit the price in reality, the R-squared value and mean absolute percentage error value are used. Firstly, the stationarity had to be checked, which is carried out with the Dicky-Fuller Test. The small p-value showed that the data don't have a unit root. Then, the data were processed by the two models written by python. According to the analysis, the multi-linear regression and random forest models are relatively accurate. However, with a larger data set or with a relatively small rolling period, the random forest will be able to achieve a better result. However, larger data comes with the price of more computing resources. These results indicate that it is feasible to predict stock market prices with machine learning models and the accuracy could be improved by utilizing better machine learning models.

Keywords: TMT Industry; Stock Price Prediction; Machine Learning; Multi-Linear Regression Model; Random Forest Regression.

1. Introduction

TMT stands for Technology, Media, and Telecommunications. Currently, it is becoming a sector that is growing rapidly and has drawn lots of investment from PE firms, venture capitalists, investment bankers, etc. Lots of TMT stocks have a huge potential to outperform the market, while the investors also have to face greater risks. 2022 is a tough year for the TMT sector, thanks to the macroeconomic headwinds, pandemic, and weak IPO market. Through the first half of 2022, the total deal value fell 64% from the first half of 2021, while the volume decreased by 21%. However, the TMT sector is looking forward to a rapid bounce back in 2023 [1]. According to the prediction of Deloitte Global, the total value of the TMT sector could increase by 25% to 50% year over year to the range of US$250-300 billion [1]. The AAPL and NFLX stock belongs to TMT, which is widely acknowledged. And because of the similarity between TSLA stock and the stocks of the tech sector, TSLA stock is included in this article [2].

There is already a lot of research concerning predicting stock market prices with machine-learning models, with the popularity rising. One article uses the model ARIMA (as known as Autoregressive Integrated Moving Average) to predict the Nokia Stock Index [3]. This model eventually achieved an adjusted R-squared value of 0.997250, which is an accurate short-term prediction [3]. In another article, a back-propagation network is utilized [4]. Three layers construct this model, including the input, hidden, and output layers [4]. The algorithm was used to predict the Gree Electric stock (a Chinese stock) and was able to get a 90% accuracy with a predicting range of 15 days [4]. Another method was composed of convolutional neural networks (CNN), attention mechanism (AM), and bi-directional long-short-term Memory (BiLSTM). These models are used to forecast the next day's closing price and this paper did a comparison of the performances of different models (MPL, CNN, CNN-LSTM, CNN-BiLSTM-AM, BiLSTM-AM, etc.) [5]. The paper reached the result that CNN-BiLSTM-AM performed the best with an R-squared of 0.9804, MAE of 21.952, and RMSE of 31.694 [5]. The ANN model and random forest model are used in another paper [6]. This paper uses the stock

price data of Nike, inc., J.P. Morgan & Co., Johnson & Johnson, etc., and reached the conclusion that ANN could be more accurate than the random forest model in that given circumstance [6].

2. Data & Method

2.1. Data

Apple Inc. is a multinational company in the technology sector and it is the largest company by both sales and market value as of May 2022, achieving sales of $378.7 billion and meanwhile enjoying a market value of $2.6 trillion [1]. Fig. 1 is the graphical representation of the AAPL closing stock price. Netflix, Inc. is an American media company that specializes in subscription video streaming service and production, reaching the rank of 115th on the Fortune 500 [2]. It is ranked 219th according to the Forbes Global 2000 [3]. Fig. 2 is the graphical representation of the NFLX closing stock price. Tesla, Inc. is an American multinational automotive, AI, and clean energy company and it is currently the 7th largest company 7 by market cap [4]. Fig 3. is the graphical representation of the TSLA closing stock price. All the financial data is from yfinance. All of them covered the period from 9th February 2018 to 9th February 2023. The first four year's data is used for training the model, and the rest is dedicated to measuring the accuracy.

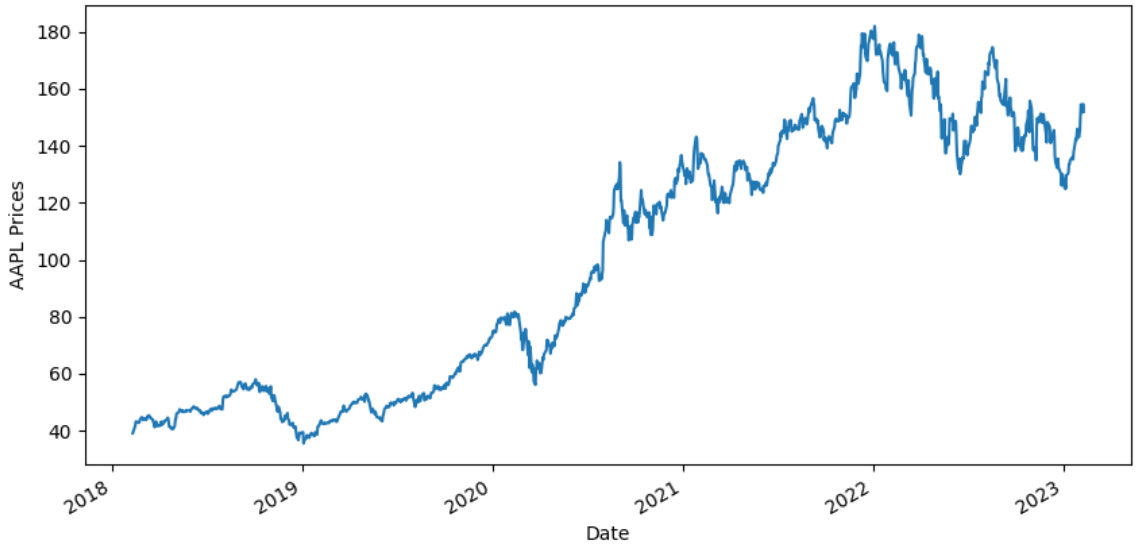

Fig 1. The AAPL stock price.

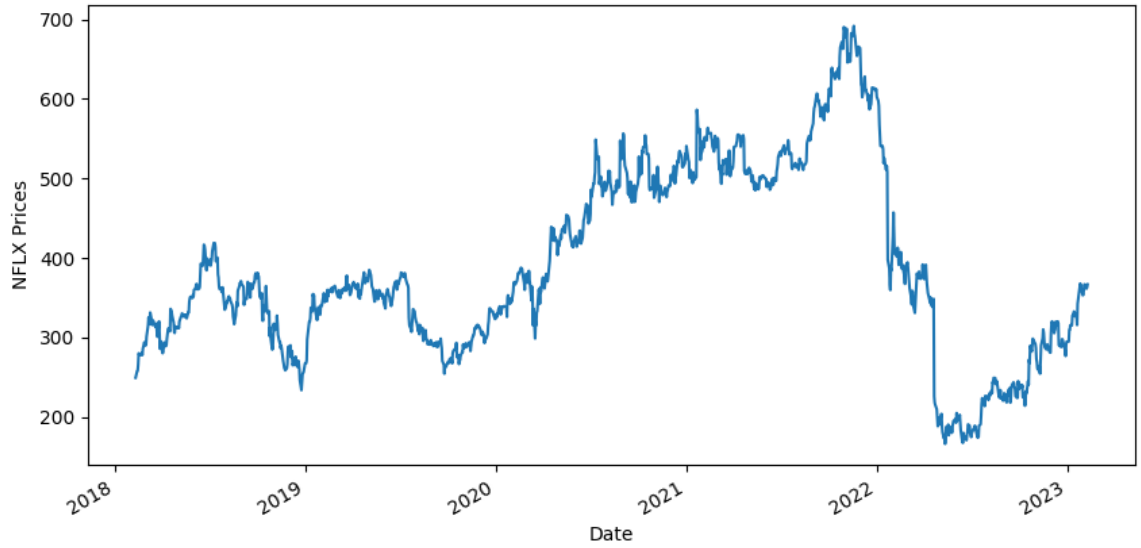

Fig 2. The NFLX stock price.

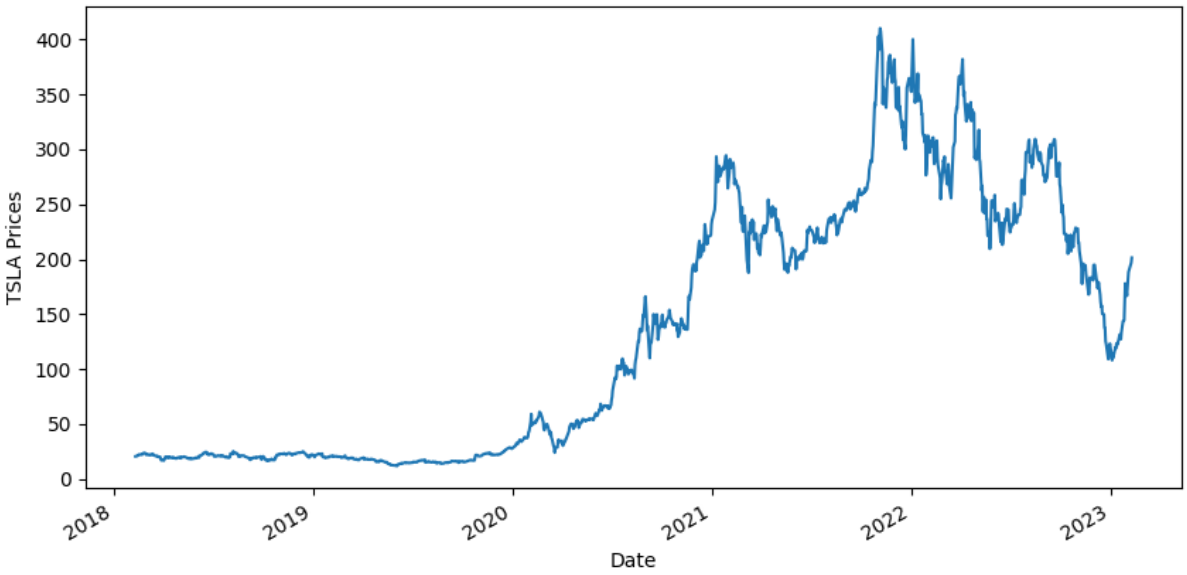

Fig 3. The TSLA stock price

2.2. Models

There will be two models used in this article: the multi-linear regression and the random forest model. The multi-linear regression model is an extension of the OLS model. OLS model assumes that the analysis fits a model of the mapping, or function between explanatory variables and a continuous outcome variable and tries to reduce the sum of square errors as far as possible [5]. The multi-linear regression model also utilizes that. Suppose there are n observations $\{x_i, y_i\}_{i=1}^{n}$ forming a data set, and for each observation i there is a corresponding scalar value y_i and a column vector x_i, i.e., $x_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T$. The response variable, y_i, is a linear function of the regressors in a multi-linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_0 \quad (1)$$

where for $i = n$ observations, $\beta_0 = y - intercept$; $\beta_p = the\ n^{th}\ slope\ coefficients$; and $\epsilon_0 = the\ residuals$. Random forest is an algorithm trademarked by Leo Breiman and Adele Cutler. This algorithm reaches a single result by combining the outputs of multiple decision trees [6]. The random forest is constructed of randomly selected data and characteristics. Random forest model can be implemented with two main parameters: the decision tree number and the feature number of the node that can be selected by each node of the decision tree.

2.3. Evaluations

The Coefficient of Determination is always denoted R^2. The following formula is used to evaluate the coefficient of determination.:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (2)$$

RSS is known as the abbreviation of the sum of squares of residues, and TSS is the total sum of squares. MAPE stands for Mean Absolute Percentage Error, and it is also known as mean absolute percentage deviation (MAPD). MAPE is a measure to express the accuracy of a forecasting method. The following formula is often used.

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad (3)$$

Here A_t is the value in reality, meanwhile, F_t is the value predicted by the models.

3. Results & Discussion

3.1. Model Performances

The data must be checked to be stationary, meaning the mean and standard deviation are not time related. It is significant because it allows for stability and certainty in the forecasting models. A null hypothesis is that a time series model has a unit root. In fact, it has to be checked by the test called Dickey-Fuller Test. Firstly, the rolling mean and standard deviation must be calculated, which is represented in Fig. 4. Dickey-Fuller Test generated the results which are represented in Table 1. With a p-value of 0.627867, the data could be presumed to be stationary. The data presented is the TSLA stock and similarly, the other two stocks also meet the requirement.

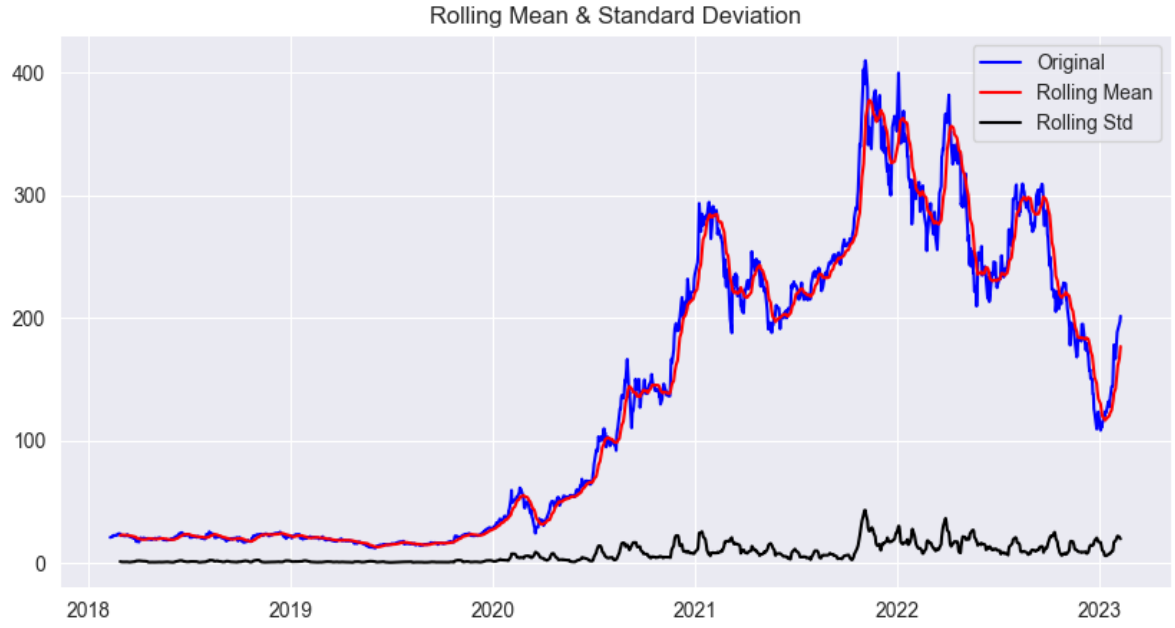

Fig 4. Rolling Mean & Standard Deviation

Table 1. Results of the Dickey-Fuller Test

Test Statistic	-1.302851
p-value	0.627867
#Lags Used	18.000000
Number of Observations Used	1239.000000
Critical Value (1%)	-3.435639
Critical Value (5%)	-2.863876
Critical Value (10%)	-2.568013
dtype: float64	

Firstly, the period of rolling is set to 3 days. The R-squared values and MAPE values of the result of the linear regression model are shown in Table 2. The random forest model used in this article has a number of trees of 1000 and a max depth of 10. The R-squared values and MAPE results are presented in Table 3. Then, the rolling period is set to 1 day. Table 4, and 5 shows the results.

Table 2. Evaluation of the Multi-linear Regression Model

Stock	R-squared Value	MAPE Value
AAPL	0.8998	2.19%
NFLX	0.9554	3.81%
TSLA	0.9592	4.38%

Table 3. Evaluation of the Random Forest Model

Stock	R-squared Value	MAPE Value
AAPL	0.9962	2.00%
NFLX	0.9743	3.21%
TSLA	0.9897	5.23%

Table 4. Evaluation of the Multi-linear Regression Model

Stock	R-squared Value	MAPE Value
AAPL	0.9311	1.75%
NFLX	0.9724	2.89%
TSLA	0.9739	3.25%

Table 5. Evaluation of the Random Forest Model

Stock	R-squared Value	MAPE Value
AAPL	0.9971	1.71%
NFLX	0.9909	2.34%
TSLA	0.9951	3.58%

3.2. Comparison & Suggestions

According to the evaluation, the random forest model tends to achieve a better R-squared value overall. However, the multi-linear regression model slightly outperformed the random forest model when it concerns MAPE value. It is widely acknowledged that the random forest model is a more sophisticated model, which takes more parameters than a linear regression model. Nevertheless, there are some cases where a multi-linear regression model can perform better than a random forest model, and linear function is one of the cases. To increase the accuracy, the increment of sampling frequency is suggested, but it will also come with the price of more data and a longer processing time. To forecast the stock market in a short term, both of the models are approved to be reliable and relatively accurate. Generally, a multi-linear regression model will perform better when the explanatory variables outnumber the noise variables. Whereas the random forest model is better when the volume of explanatory variables increases. Therefore, the use of the random forest model is suggested.

4. Limitations & Future Outlooks

In this article, although the fitting is relatively accurate, the study has its limitations. Firstly, the sample set is limited. The study only used three stocks to represent the TMT sector. Stocks in the TMT sector have a huge number of influencing factors, such as political events, inflation, interest rates, etc. Therefore, different stocks may have significant differences in terms of characters, which will have a big influence on the accuracy of the model fitting. The relation between predicting accuracy and some influencing factors can be explored further. Meanwhile, the model used in this article is not very complex. In future research, more complex models or combinations of multiple models could be utilized to create a more powerful and accurate model. Additionally, the performance difference between multiple models (CNN, LSTM, random forest, etc.) can be tested. Additionally, the models could use other inputs other than the prices of the past. For instance, the relationships between the stock prices of multiple companies can be explored. Some companies share mutual interests and others have direct interest conflicts, which might be possible to present by the correlation between the stock prices.

5. Conclusion

In summary, this study used a multi-linear regression model and random forest model to predict the closing stock price of Apple Inc., Netflix, Inc., and Tesla, Inc. The result of the fitting is relatively accurate. According to the analysis, both the multi-linear regression model and the random forest

model are reliable in this circumstance. It also shows that less rolling period could result in a more accurate fitting and random forest is better than a multi-linear regression model in most cases. Nevertheless, the result is not without its limitations. The complexity of the models and the number of stocks used in this article need future improvement and the factors behind the behavior of the stock prices could be explored. Apart from the limitations, this research shows that it is feasible to forecast the short-term stock price with machine learning models and that the multi-linear regression model and the random forest model are reliable in this circumstance.

References

[1] Technology, Media, and Telecommunications Predictions 2023. Retrieved from: Deloitte Insights, www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions.html#explore. Accessed 13 Mar. 2023.

[2] Sharma R. Tesla: Tech Company or Car Company? Investopedia, 25 June 2019, Retrieved from: www.investopedia.com/news/tesla-tech-company-or-car-company/.

[3] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model. 2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014: 106-112.

[4] Zhang D, Sha L. The Application Research of Neural Network and BP Algorithm in Stock Price Pattern Classification and Prediction." Future Generation Computer Systems, 2021, 115(1): 872–879.

[5] Lu W, Li J, Wang J, et al. A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 2021, 33: 4741-4753.

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

[7] Ponciano J. The World's Largest Tech Companies in 2022: Apple Still Dominates as Brutal Market Selloff Wipes Trillions in Market Value. Forbes, Retrieved from: www.forbes.com/sites/jonathanponciano/2022/05/12/the-worlds-largest-technology-companies-in-2022-apple-still-dominates-as-brutal-market-selloff-wipes-trillions-in-market-value/?sh=4e8ee8393448. Accessed 9 Feb. 2023.

[8] Netflix | 2020 Fortune 500. Fortune, Retrieved from: fortune.com/company/netflix/fortune500/.

[9] Netflix | NFLX Stock Price, Company Overview & News. Forbes, Retrieved from: www.forbes.com/companies/netflix/?sh=7a360b0d8541. Accessed 9 Feb. 2023.

[10] Companies Ranked by Market Cap - CompaniesMarketCap.com. Companiesmarketcap.com, companiesmarketcap.com. Accessed 12 Feb. 2023

[11] Zdaniuk B. Ordinary Least-Squares (OLS) Model. In: Michalos, A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht 2014.

[12] Ma Y, Han R, Fu X. Stock prediction based on random forest and LSTM neural network. 2019 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea (South), 2019: 126-130.