Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/45/20230269

Using Machine Learning Models to Predict the Uber Stock

Jingyu Gao[1,a,*]

[1]*Dalian University of Technology, Panjin Liaoning 124000, China*
a. 853075582@dlut.edu.cn
**corresponding author*

Abstract: This paper aims to describe how to use machine learning models to predict the situation changing of stocks. This paper will use linear regression and random forest model to predict the Uber stocks stock's future closing price and probability of rise and fall. This paper firstly collected stock related information from Kaggle. The data of Uber stocks are from May 10, 2019 to March 24, 2022. The closing price and the future closing price are divided by taking 80% as the training set and 20% as the proportion of the test set. Then setting some technical indicators to analyze the accuracy and deviation of the prediction, such as root mean square error (RMSE), mean deviation error (MBE) and R-square. In future research, these methods could be used to apply machine learning models in stock forecasting, as well as other more accurate methods such as radio frequency technology and neural networks.

Keywords: linear regression, random forest, stock prediction

1. Introduction

The idea behind this software is that it can make transportation easier and faster in our daily lives, which has been put forward by Garrett Camp, who is the co-founder of StumbleUpon [1]. After the beta is available on the public application in May 2010, In San Francisco, Uber's mobile application publicly launched in the next year [2]. In May 2019, Uber is listed on the New York Stock Exchange [3].

Afterwards, Uber suffered a period of world public health battering. Focusing on the most recent quarter of the 2023, Uber posted its strongest quarterly report ever, as a company that survived public health events by diversifying its business is now reaping the benefits of a post-pandemic era. The Uber's ride-hailing business had been hit hardly by the coronavirus pandemic. However, it did bounce back to the levels which is almost same as the pre-pandemic in Q4 2020. During this time, Uber food delivery has played a significant role in its business. By 2022, the revenue of ride-hailing had surpassed the food delivery and become the driver of operating profit.

DOI: 10.54254/2754-1169/45/20230269

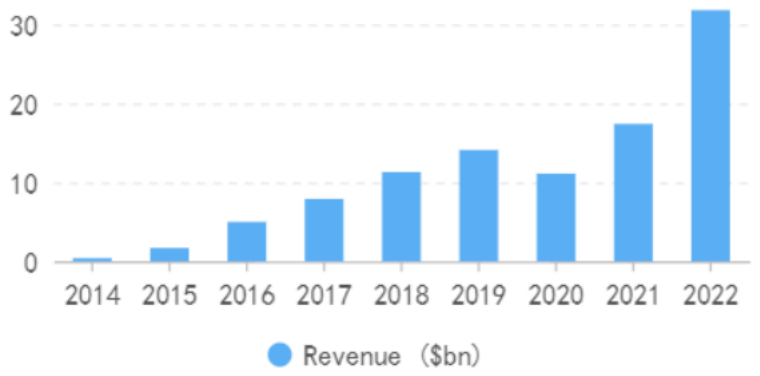

Figure 1: Uber's annual revenue ranged from 2014 to 2022. (Photo credit: Original).

Figure 1 shows the Uber annual revenue 2014 to 2022. In 2022, Uber made about $31.80, up 82 percent from last year. However, Uber's revenue showed a significant decline due to the global pandemic of the novel coronavirus in 2020. [4].

All three of Uber's major divisions showed rising revenue in 2022. Freight revenue remained the highest, up 22.8 percent from the previous year, higher than revenue from transportation and delivery. [4].

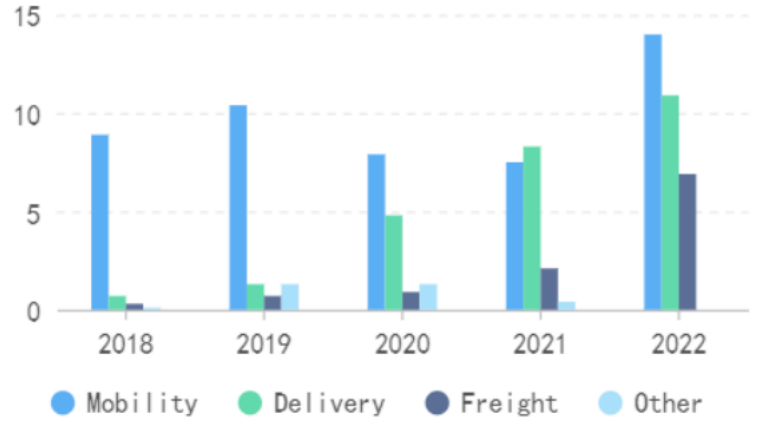

Figure 2: The average annual revenue of Uber's divisions runs from 2018 to 2021. (Photo credit: Original).

Figure 2 indicates that the Uber annual revenue by segment 2018 to 2021.From the figure 2, Uber has three main divisions--freight, transportation and distribution, all reported higher annual revenue between 2021 and 2022. Freight has the highest income in 2022. [4].

The dynamics and uncertainty of the stock market make predicting stock prices a difficult task, because it depends on different internal and external factors. Some external factors such as trading conditions and economic factors [5,6].

There has been some research on stock forecasting before. Classically, there are two main ways to forecast stocks. One analysis method is qualitative analysis of stocks, which is based on most external factors different from stock data, like the company's development, trading conditions and political

DOI: 10.54254/2754-1169/45/20230269

factors. Another kind of technical analysis uses historical prices, for instance, closing price and opening price, to predict the price of a stock quantitatively. Machine learning already plays an important role in stock price predicting. In particular, the prediction of stocks with the huge non-linear data scale, these things are not predicted by the previous methods. In order to properly analyze these data, the complex relationships and hidden relationships they contain need to be solved by thinking such as machine learning [7]. In most computing fields, some classical algorithms are used, such as linear regression [8], random walk theory (RWT) [9], moving average convergence/divergence (MACD) [10]. The research has shown that using machine learning can improve the accuracy of stock market predictions. These include techniques such as support vector machines (SVM). Some artificial neural networks (ANN), convolutional neural networks (CNN), recursive neural networks (RNN) and deep neural networks also show good results.

Breiman proposed random forest, and the randomness was enhanced on the basis of bagging. In addition to the data that each tree lists, random forests also change the way they are constructed. In the case of a standard tree, each node optimally splits all variables. This strategy performs very well compared to many other classifiers [5].

Therefore, to predict the stock, collecting Uber stock data information on kaggle, which can be obtained a series of contents such as the opening price, the highest price and the lowest price. Then the mean value, the maximum value, the minimum value and the variance of daily data can be analyzed, because it is conducive to having a deep understanding of the data. Finally, the data is normalized (which is not done in the Random Forest model), and the above is the preliminary data preparation process. The linear regression model mainly calculates the explanatory variables and the explained variables through its own mathematical model, so as to obtain the calculation results. Random forest is the main step of random forest classifier, so as to predict the probability of stock rise and fall.

The rest of this article is organized in the following sections. In Section 2, we describe the methodology, their derived variable, and how they were processed into the model. In the following Section 3, we present and describe the results and compare random forest with linear regression by some technical index, such as MBE, RMSE and R^2. Finally, we conclude by describing our future development.

2. Methodology

2.1. Random Forest

A random forest is a good approximation. Because of the huge size of the stock, this can lead to a high level of noise in the stock, which usually affects the price of the market. Trees grow in a completely different way than expected. Its purpose is to minimize prediction error by processing.

Analyzing the Data: Through the stock data of Kaggle, the data from May 10, 2019 to March 24, 2022 are collected, and the average calculation function in machine learning is used to obtain the descriptive data statistics of two important indicators in the stock -- the opening price and the closing price. Descriptive statistics of data (sample size, mean value, standard deviation, maximum value, minimum value). All the stock data are from the Kaggle (Data source:https://www.kaggle.com/datasets/varpit94/uber-stock-data).

DOI: 10.54254/2754-1169/45/20230269

Table 1: Data Analyzing.

Data set	Open	Close
Mean	40.16	40.11
Max	63.25	63.18
Min	15.96	14.82
Std	9.20	9.14

Table 1 reports some basic statistical characteristics of the mean, max, min, and standard deviation of the opening price and the close price, and it may be seen that Uber's stock price will be stable in the $40 range. The standard deviation of the data reflects the degree of dispersion. The standard deviation of the opening price is greater than that of the closing price, which might indicate that the stability of the opening price is higher than that of the chassis price. However, their standard deviation is relatively high, which also reflects the degree of difficulty in predicting the price.

Derived variable: Four new variables were created for predicting stock closing prices. These four variables were used to assess the importance to the forecast. There are four variables, such as close-open, high-low, MA5 (Five-day average) and MA10 (Ten-day average) were used to have a clearer understanding of the stock data in macro aspects, such as knowing the general trend of the stock's rise and fall, and the average closing price of the stock within 5-10 days.

Close-open indicates that using the closing price minus the opening price, which shows the stock’s final performance for the day. A positive number means the stock is rising, a zero number means the stock is stable, and a negative number means the stock is falling.

High-low indicates that using the highest prices minus the lowest price, which refers to the oscillation range between the lowest price and the highest price of Uber stock in a day. It might reflect the degree of stock activity to a certain extent

MA5 means the average of a stock’s closing price over five days. It is a short-term moving average in the moving average system, which basically maps the short-term movement of the stock closing price, same as the MA10.

Model Setup: The training set and the test set are divided by taking 80% as the training set and 20% as the proportion of the test set. However, it is divided according to time series, not by train test split function, because stocks have time.

2.2. Linear Regression

Through linear regression model to predict the stock trend and rise and fall.

$$y=kx+b \tag{1}$$

This equation reveals the mathematical model of linear regression. In this equation, the explained variable y represents the closing price of the stock 40 days after the date, and x represents the closing price of the stock on the day. The k and b are both parameters of this equation and are obtained based on the calculation of the closing price of Uber stock by the linear regression algorithm. Finally, through the construction of the time series model, each variable x corresponds to a specific date, so this equation can predict the closing price 40 days after the known dates.

Derived Variable: A new variable "target" has been created for predicting stock closing prices, which has been used to train the model. The "target" variable is actually the closing price 40 days later.

Model Setup: The training set and the test set are divided by taking 80% as the training set and 20% as the proportion of the test set. First, the linear regression is initialized with the name of the model

DOI: 10.54254/2754-1169/45/20230269

and used as the fitting object. At the same time, the accurate fitting value is calculated to reflect the performance of the model. Finally, the value is predicted and the closing price in the next 40 days is successfully predicted.

3. Results

To evaluate the validity of these models, linear regression and random forest were compared. The predicted closing price is affected by root mean square error (RMSE) and mean deviation error (MBE) to find the final minimum error of the predicted price.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Oi-Fi)^2}{n}} \tag{2}$$

Where i refers to the initial closing price, Fi refers to the predicted closing price, and n refers to the total window size.

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(Oi - Fi) \tag{3}$$

Where i refers to the initial closing price, Fi refers to the predicted closing price, and n refers to the total window size. Figure 2 shows the use of artificial neural networks.

Table 2: Technical index.

Data set	Random Forest	Linear Regression
MBE	1.64	6.97
RMSE	1.28	2.64
R2	-0.65	0.10

Table 2 shows that the technical index in different models. The first line reports the mean deviation error of the random forest and linear regression, also the same as RMSE and R^2. From table 2, linear regression is better than random forest in R^2.

However, random forest is superior to linear regression in terms of RMSE and MBE.

Table 3: The probability of the price change.

	Random Forest	Linear Regression
Up	0.54	0.51
Down	0.46	0.49

Table 3 shows the prediction of random forest model and linear regression model on the probability of stock rise and fall. The data analyzed are all from the data predicted by the models. Linear regression predicts probability by diverting the forecast from the closing price of the previous day. If the difference is greater than 0, it is regarded as an increase; if the difference is less than 0, it is regarded as a decline. The random forest model, however, has its own classifier to analyze the probability of a stock's rise or fall. The first line reports the probability that the stock price rises, the second line reports the situation of downing. From table 3, it means that different models predict different results. The probability of stock rising in random forest predicting is higher than that in linear regression.

DOI: 10.54254/2754-1169/45/20230269

Table 4: The accuracy.

	Random Forest	Linear Regression
Score	0.59	0.48

The table 4 shows the accuracy between the different models. The accuracy is calculated by using the scoring model to calculate the training set and the test set of the random forest model and the linear regression model respectively. From Table 4, the accuracy of random forest is greater than that of linear regression.

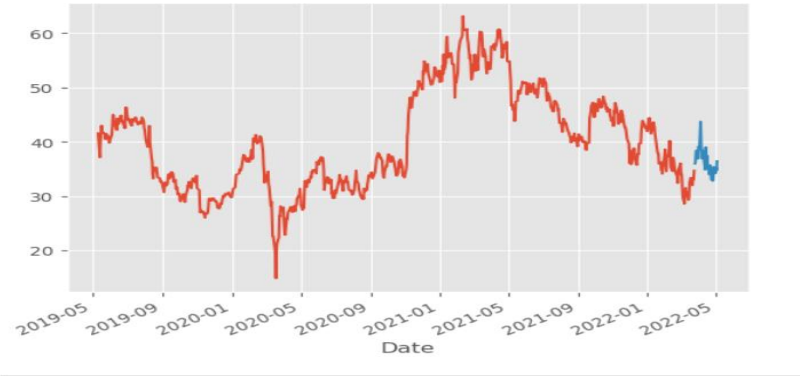

Figure 3: The fitting plot of Linear Regression.

Where the red line refers to the true price, which shows the original closing price from May 10, 2019 to March 24, 2022. The blue line refers to the predicted line, which shows the closing price predicted by the linear regression. The horizontal axis is the time series, and the longitudinal axis is the closing price. Figure 3 is based on the original closing price of the stock, then the predicted closing price of the stock is overlaid to get the curve. It shows how much the closing price has gone up or down, with the largest range ranging from $14.82 to $64.81.

Comparing random forest with linear regression, random forest collation proved to be a better technique, providing better RMSE, MBE values and accuracy, as shown in Table 2 and 4. As revealed in Table 2, for linear regression, this study only predicted the stock closing price after 40 days. Therefore, the RMSE and MBE of linear regression model might be affected by the predicted days. It is believed that 40 days is not the best predicted days, which is quite different from the RMSE and MBE predicted by random forest classifier. Meanwhile, from the three data, the fitting accuracy of random forest model is obviously higher than that of linear regression.

4. Conclusions

Forecasting the stock market is a challenging task in that stocks depend on multiple parameters to form complex forms. In conclusion, the main problem of this study is to predict the closing price of Uber stock through linear regression model and random forest model. Through the study, it also successfully predicted the probability of the stock's rise and fall on the next day and the closing price in the future.

The idea of this study is to calculate and fit the closing price of stocks in the next 40 days through the mathematical model of linear regression, and make statistics on the rise and fall of stocks through the data of the closing price of stocks. The idea of random forest model is to analyze the rise and fall of stocks through its own classifier.

DOI: 10.54254/2754-1169/45/20230269

Finally, comparative analysis based on RMSE, R^2 and MBE values clearly shows that random forest is a better predictor of stock prices than linear regression, and the fitting accuracy of random forest model is obviously higher than that of linear regression.

For future study, the linear regression model could be improved by using more data (not only 40 days), which might improve the accuracy of forecasts. In data processing, it also could eliminate some relatively unusual points, such as daily limit up and daily limit, and reduce the R^2 of the model. However, there are some limitations, like the value of the max depth is not appropriate that might lead the over fitting, the estimators might affect the decision trees. In order to enhance the prediction accuracy of the model, it can be developed in the stock prediction, such as radio frequency technology and neural network.

References

[1] *Scott, A.: Co-founding Uber made Calgary-born Garrett Camp a billionaire. Canadian Business (2015).*

[2] *Lagorio-Chafkin, C.: How Uber is going to hire 1,000 people this year. Inc. Archived from the original on November 18, 2018.*

[3] *Driebusch, C., Farrell, M.: Uber IPO stumbles, stock trades below offering price. The Wall Street Journal (2019).*

[4] *Mansoor, I.: Uber revenue and usage statistics. Business of Apps, 2023.*

[5] *Vijh, M., Chandola, D., Tikkiwal, V. A., Kumar, A.: Stock closing price prediction using machine learning techniques. Procedia Computer Science 167, 599-606 (2020).*

[6] *Hur, J., Raj, M., Riyanto, Y.E.: Finance and trade: A cross-country empirical analysis on the impact of financial development and asset tangibility on international trade. World Development 34(10), 1728-1741 (2006).*

[7] *Li, L., Wu, Y., Ou, Y., Li, Q., Zhou, Y., Chen. D.: Research on machine learning algorithms and feature extraction for time series. IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 1-5 (2017).*

[8] *Seber, G.A.F., Alan, J.L.: Linear regression analysis. John Wiley & Sons 329 (2012).*

[9] *Reichek, N., Devereux, R.: Reliable estimation of peak left ventricular systolic pressure by M-mode echographic-determined end-diastolic relative wall thickness: Identification of severe valvular aortic stenosis in adult patients. American Heart Journal 103(2), 202-209 (1982).*

[10] *Chong, T. T., Ng. W.: Technical analysis and the London stock exchange: Testing the MACD and RSI rules using the FT30. Applied Economics Letters 15(14), 1111-1114 (2008).*