

Stock Price predication based on Linear Regression, RNN, LSTM

Bowen Xia^{1, *}

¹Department of Mathematics & Business Administration, D'Amore-Mckim School of Business, Northeastern University, Boston, United States

*Corresponding author: xia.bow@northeastern.edu

Abstract. The prediction of stock price has always been a topic of concern by the financial analysis industry. With the progress of science and technology, the prediction methods of stock price are becoming more and more rich and advanced. From classical linear regression to the state-of-art neural network scenarios, from mathematics to AI deep learning, the introduction of computers and the development of programming also provide a new direction for the prediction of stock prices. This study takes the stock of Johnson & Johnson in the past five years as an example, selects the open, high, low price of the stock as the explanatory variable, and chooses the close price as the explanatory variable to establish a linear regression model. This study also establishes RNN and LSTM models to predict stock prices. Subsequently, the R^2 , RMSE, MAE of the model are evaluated, where LSTM model have the best R^2 score and the best prediction of future stock price. These results shed light on guiding further exploration to the short-term stock price prediction.

Keywords: Stock Price Prediction; Machine Learning; Linear Regression; Random Forest; LSTM.

1. Introduction

Stock price trend has always been a topic of concern for stock investors, financial analysts, enterprise evaluators, economists etc. The stock price determines the development of a listed company and can reflect the current situation of the industry in which it is located. The way to predict the future price of stocks more accurately has always been the goal of researchers. Stock is one of the daily financial management methods of some people. The change of stock price is also an issue of concern to the people and the country. The stock prices in the stock market are sequence dependent, which means that the historical information of stocks may be used to predict future stock prices.

According to previous study, there is a certain linear relationship between stock prices [1]. In this paper, the historical data of stocks are input into SPSS software, which is utilized to process the data to get the R^2 value of the model. Besides, DW value is used to judge that there is no serial correlation in the model. Finally, the prediction effect of the model is judged by the error rate, and it is concluded that the multivariate linear model can predict the stock trend accurately in the short term.

Early in 1998, there was research on stock price prediction based on neural network [2]. In this study, it was mentioned that researchers studied how to use prior knowledge and neural network to improve multivariate prediction ability. The study uses the previous stock price prediction model and domestic and foreign news. In addition, through previous studies, this study selects several economic indicators and events to input them into the neural network. Finally, it is concluded that at the 5% significance level, the prediction error of the researcher's method is less than that of multiple regression analysis.

Moreover, other researchers shows that it is difficult to predict the stock price trend, mainly because most of the factors affecting the stock price are complex, incomplete, and unclear [3]. In this study, the price data of financial products are regarded as a one-dimensional sequence generated by projecting the chaotic system composed of multiple factors into the time dimension, and the phase space reconstruction (PSR) method of time series is used to reconstruct the price sequence. Based on the linear programming method and the long-term and short-term memory network (LSTM), a prediction model based on DNN is designed to predict the stock price.

Based on the above research, this study selected the stocks of Johnson & Johnson listed companies in the United States, built models in Python, and used multiple linear regression, RNN, and LSTM

models to predict the stock price. As a beginner of machine learning, neural network and python, the author intends to show the usage of some models through short-term study, so as to help future beginners better understand these models and leave some ideas to more professional researchers according to some findings in the research process. The rest part of the paper is organized as follows. Sec. 2 will introduce the methodology in this study. Sec. 3 is result and discussion, which will introduce the correlation between variables; Comparison of several methods; empirical analysis and limitation. Sec.4 is the summary of all.

2. Methodology

The data of this study comes from the stock price of Johnson & Johnson on Yahoo Finance. The start date is July 3, 2017, and the end date is July 7, 2022. CSV file data includes the date of Johnson & Johnson stock, open price, high price, low price, close price, adjust close price and volume. The visualization and monthly statistical results of the datasets are given in Figs. 1-3.

Monthwise comparison between Stock actual, open and close price

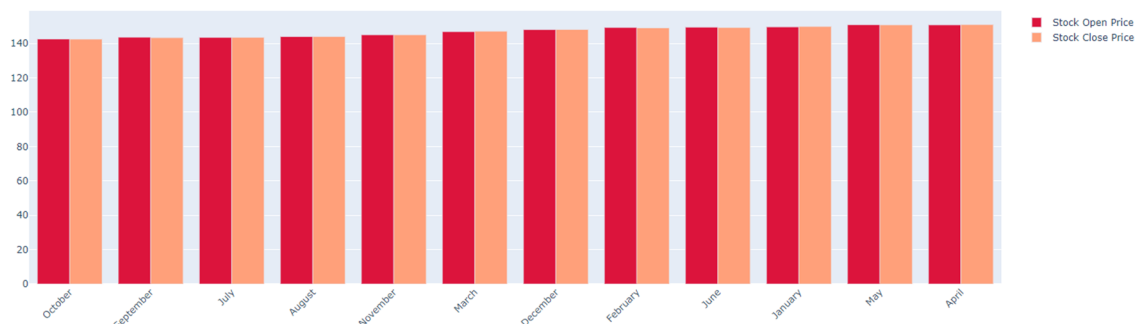


Fig. 1 Monthwise comparison between Stock actual, open and close price

Monthwise High and Low stock price

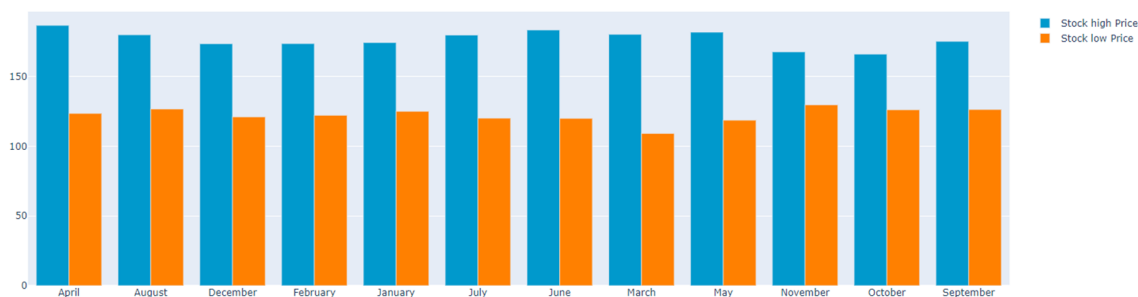


Fig. 2 Monthwise High and Low stock price

Stock analysis chart



Fig. 3 The stock price chart

The first prediction method used in this research is multivariate linear model [4]. Multivariate refers to longitudinal data modeling, which means that an individual is measured at several different time points. Another Multivariate linear regression refers to nested data modeling that multiple individuals in each cluster. Here is the formula of multivariate model:

$$Y_{n \times p} = X_{n \times (k+1)} \beta_{(k+1) \times p} + \epsilon \quad (1)$$

where Y is multiple dependent variable; X is the predictor variables; β is coefficient.

Random forest is an integrated learning method for classification, regression, and operates by building multiple decision trees at training time [5]. For regression models, a random forest returns the average and average predicted values for each tree.

Finally, this study also selects the LSTM model of RNN to predict the stock price. Long short-term memory (LSTM) is a popular RNN architecture, which was introduced by Sepp Hochreiter and Juergen Schmidhuber [6]. LSTM model can solve the problem that the basic RNN model cannot predict the current state of the model. LSTM has three "cells": input gate, output gate and forgetting gate, to help it filter and predict the information required by the neural network.

The evaluation standard of the research will be realized through the programming calculation of python, including RMSE, MAE and R^2 to determine whether the train and test sets of data meet the estimation standard and whether the model can successfully predict the future stock price. R^2 represents the variance of the model. The higher the score of R^2 , the more accurate the model is. However, if R^2 is 1 for the data of the test set, it means that the model may over fit or miss the data, or the training set coincides with the test set. Root mean square error (RMSE) can better describe the data. Since the magnitude of MSE unit is different from that of error, and RMSE is at the same level as data, which makes it easier to perceive data. The mean absolute error (MAE) is the average of the absolute values of the errors. The smaller the RMSE and MAE, the smaller the difference between the predicted and actual values of the model. All data processing in this study was performed using Google lab. Data modeling comes from Kaggle, a Python code shared by others. The stock price prediction is realized by fine-tuning the code.

3. Results & Discussion

3.1 Correlation analysis

For multivariate linear regression, the first step of experimental operation is to find the linear relationship between different variables. The research establishes the cormap between different variables through the code, and obtains the correlation between different variables and the change law between variables through the function of Python. From the Cormap (seen from Fig. 4) and the calculated results (scatter plots illustated in Fig. 5), there is a strong correlation between other variables except volume. With close price as the independent variable, open price, high price, low price, adj close price all have a strong positive correlation with them.

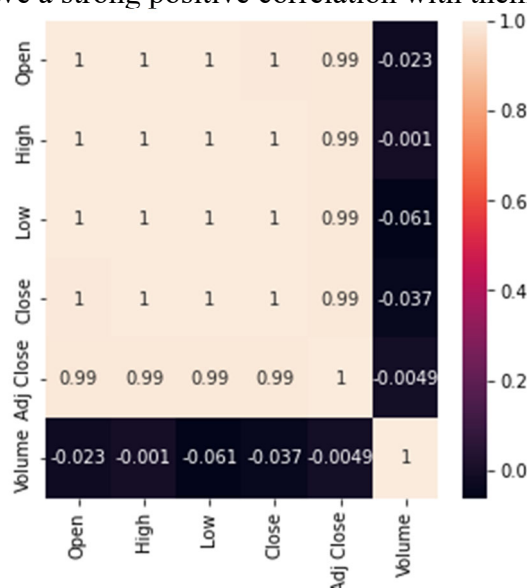


Fig. 4 The heatmap of correlation analysis.

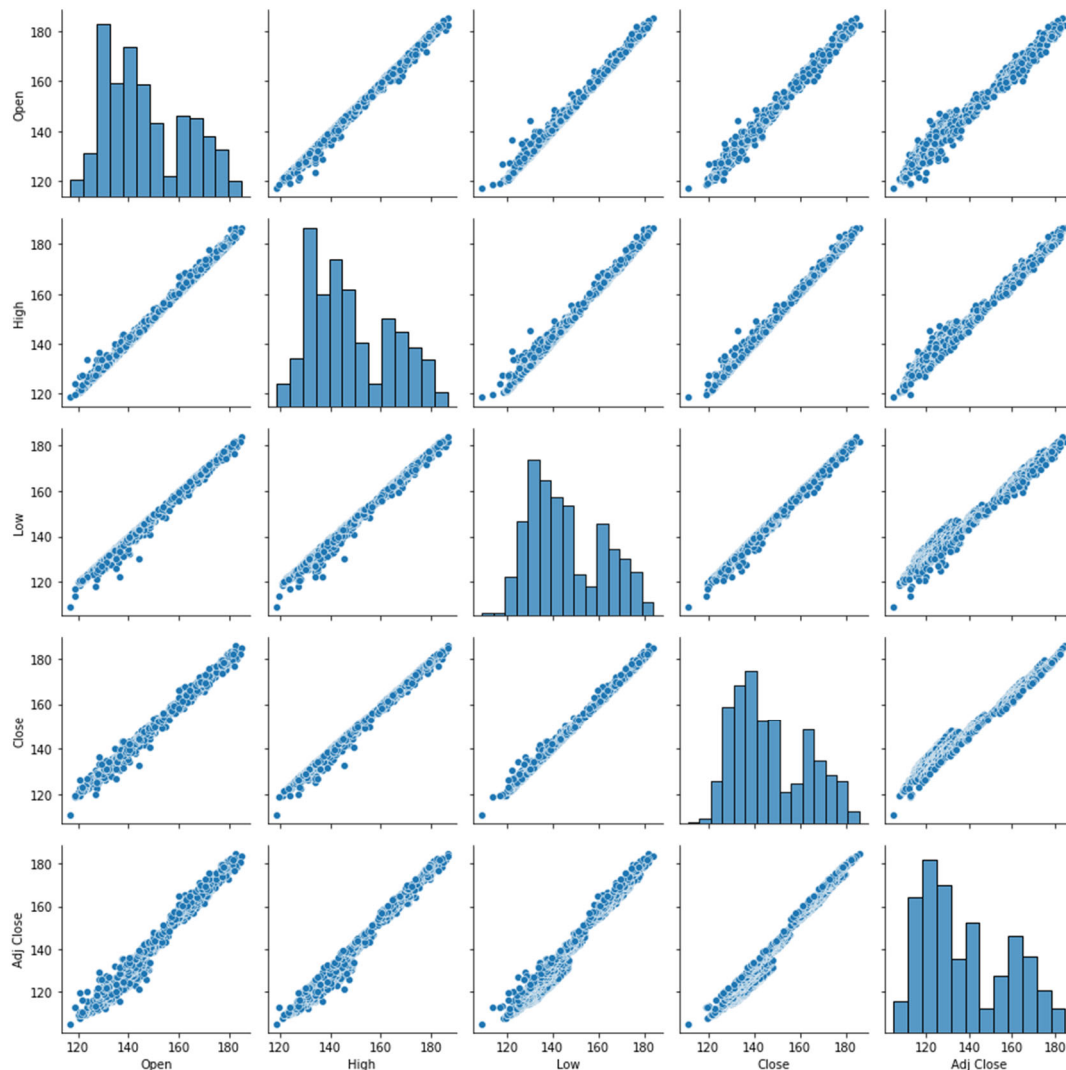


Fig. 5 Relationship scatter plots

3.2 Comparison

Since other parameters have linear relationship with close, the multiple linear regression model is chosen to predict stock price. The experiment uses linear regression in Python's sklearn to build the model. Since range of data in different columns very significantly, the experiment needs to scale the independent variable. Therefore, min max scaling is used in the experiment.

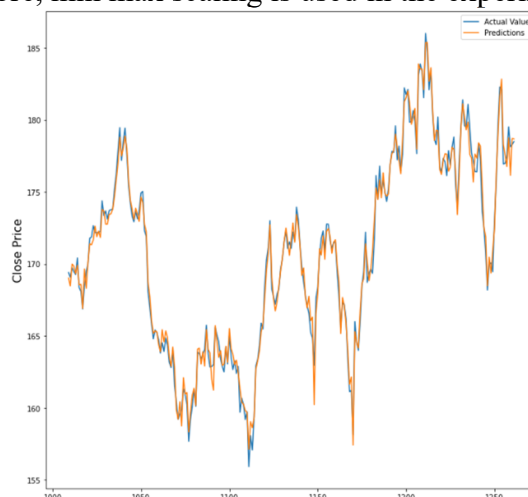


Fig. 6 Linear Regression Actual Value & Prediction

The linear regression model of research fitting has extremely high fitting data and minimal error on the train set, but when the researchers put the model into the test set, the prediction value of the model is always unsatisfactory. Besides, the stock price in the next ten days cannot meet the expectations of the research. Therefore, the R^2 , RMSE and MAE for the test set data of OLS model are not included in Table. 1. It is preliminarily concluded that it is caused by over fitting of the model. The prediction results of multivariate linear model are given in Fig. 6. where the metrics are listed in Table. 1.



Fig. 7 Random Forest R^2 Rating figure.

In the experiment random forest model, first import random forest regenerator from libraries: sklearn. The model has 100 estimators and random state=0. The data is divided into train set and test set, with a ratio of 85:15. Use scaler to transform the model back to original form. Evaluation metrics RMSE, MAR, and R^2 values are listed in Table. 1. Using the trained model, the prediction results of the stock price in the next ten days is presented in Fig. 8 and Fig. 9.

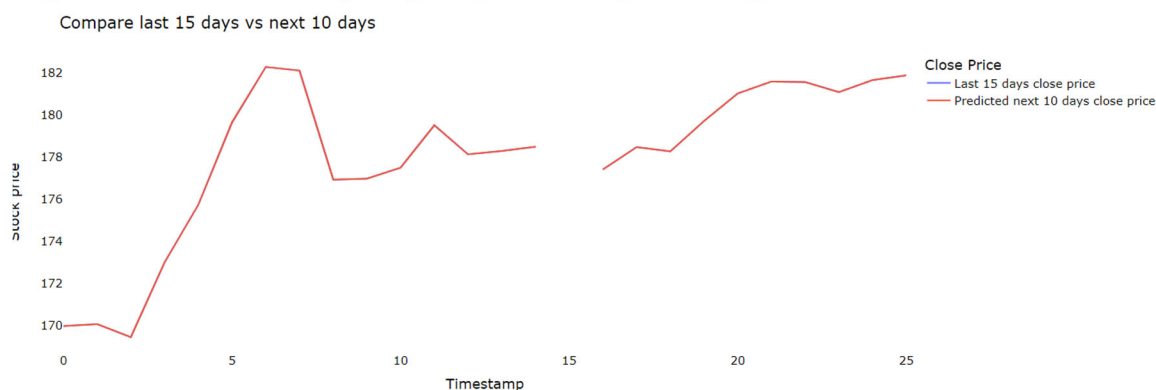


Fig. 8 Random Forest next ten days stock price prediction

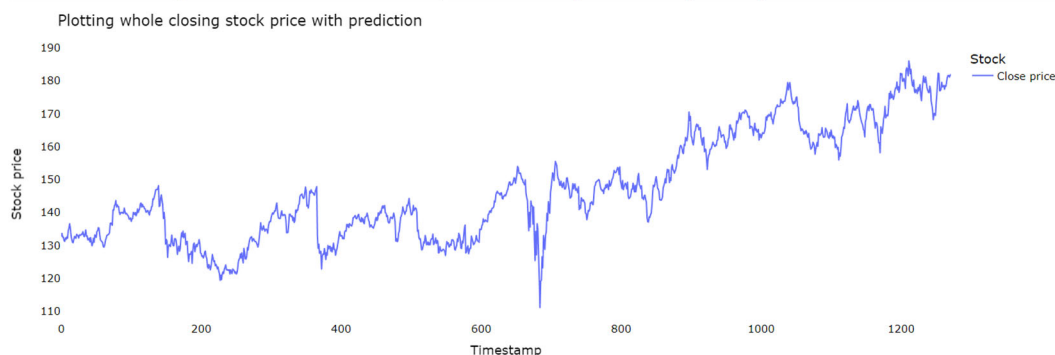


Fig. 9 Random Forest stock close price Prediction

In building the LSTM model, first reshape input to be [samples, time steps, features], which is required for LSTM. Import Tensorflow in libraries, use AdamOptimizer, and set the data to fit model, with the parameters epochs=200, batch_size=5, verbose=1. Then, one needs to check performance metrics and transform data to original form. Evaluation metrics RMSE, MAE and R^2 values are listed in Table. 1. The prediction results are demonstrated in Figs. 10-12.

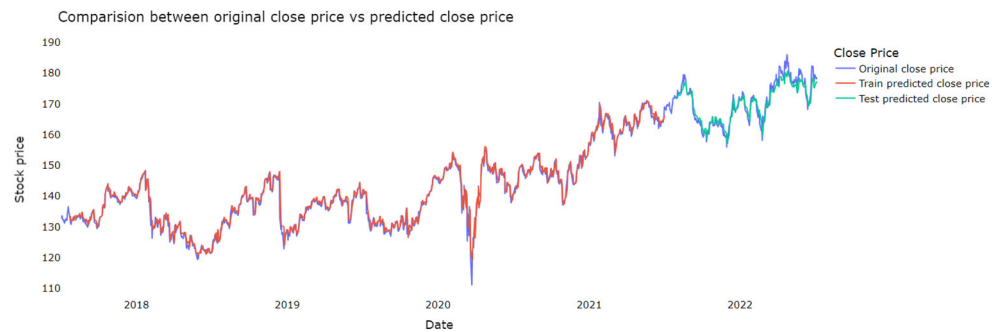


Fig. 10 LSTM R^2 Rating figure

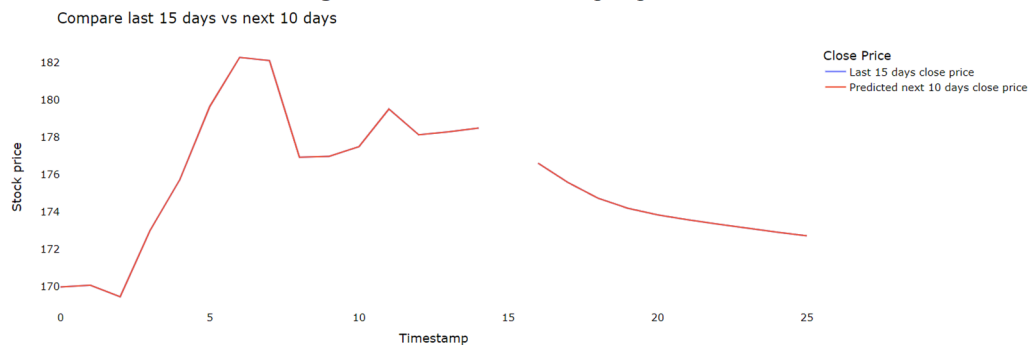


Fig. 11 LSTM next ten days stock price prediction

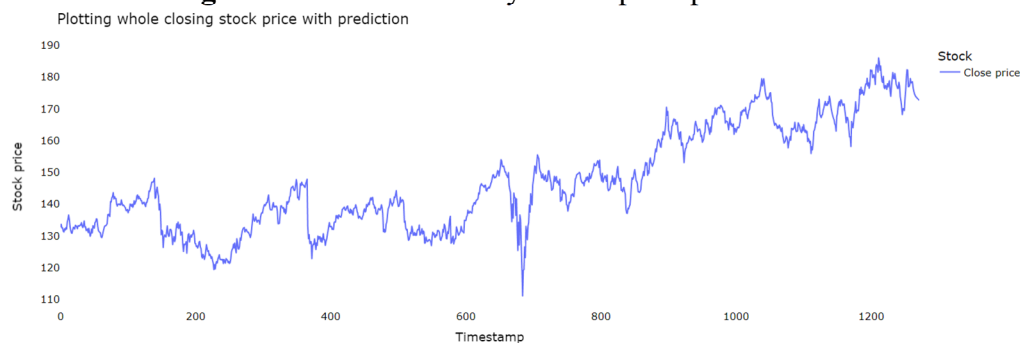


Fig. 12 LSTM whole stock close price Prediction

Table. 1 Different metrics for OLS, Random Forest, LSTM

	Train R^2	Train RMSE	Train MAE	Test R^2	Test RMSE	Test MAE
OLS	0.9887	0.71	0.56			
Random Forest	0.9970	0.73	0.48	0.8490	2.66	1.95
LSTM	0.9765	1.78	1.21	0.9039	2.12	1.64

3.3 Empirical analysis

Comparing the train set R^2 , RMSE and MAE of the three models, the fitting of the random forest regression model is better than that of OLS and LSTM. The Random Forest model has an R^2 closer to 1, which means that its model is the closest to the real situation. Meanwhile, the smaller RMSE and MAE values also show that at the training phase the Random Forest regression model has smaller errors compared with the other two models.

The results obtained from the experiment seem unexpected, but in fact they are obvious. With only open, high, low, close, adj close price as parameters, multiple linear regression is often efficient in short-term prediction, but less accuracy in testing. Random forest model has the best train set data, which represents the closest real data it fits in the three models. However, due to overfitting, the performance of random forest model in test set is not as good as that of LSTM model. Limited by the number of variables and the complexity of the model setting, the results show that random forest and LSTM models can predict the stock price well in the short term. Nevertheless, in real life, the stock price will be affected by many factors, including people's income, news, national policies, the

international situation, and even the weather. These complex uncertainties often affect the price of a stock. These are the advantages of LSTM model over multiple linear regression model. LSTM model learns the long-term dependence between variables by remembering information for a long time. The author of this paper once made an interesting simulation when studying and learning RNN model at the beginning and selected the stock price of Johnson & Johnson company for two periods to predict the future stock price. One is from July 3, 2017, to July 7, 2022 (see in Fig. 13), and the other is from July 3, 2017, to February 1, 2020 (that is, the initial stage of covid-19 in the United States [7] ,see in Fig. 14). The results are unexpected. The following are the model diagrams of two different time nodes.

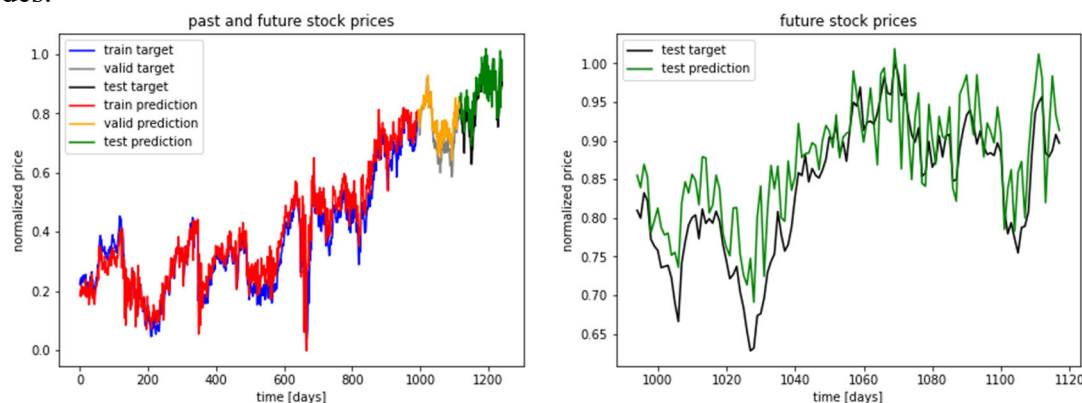


Fig. 13 RNN Prediction from 2017.7.3-2022.7.7.

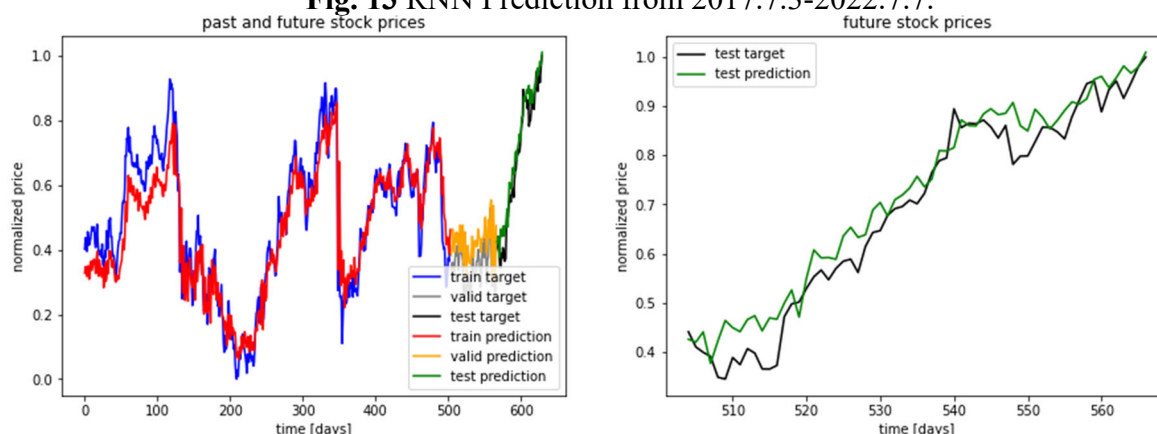


Fig. 14 RNN Prediction from 2017.7.3-2020.2.1.

Under the condition that other model parameters are completely consistent, there are two distinct results, and it is not difficult to find that the prediction of Johnson & Johnson's stock price from July 3, 2017, to February 1, 2020 is incorrect through objective facts. However, the recent model prediction is not much different from the actual stock price. It is speculated that one probable reason is that the recent health stock market is relatively stable. Therefore, the form-based prediction is obvious. COVID-19 had a serious impact on the global economy, i.e., the stock price fell. Whereas this is difficult to quantify and predict in the known data of stock prices. Nowadays, a considerable number of financial researchers are still studying the impact of COVID-19 on stock prices [8]. A study entitled "determinations and prediction of the stock market during COVID-19: evidence from Indonesia" shows that the influence of covid-19 on stock price through FFT curve fitting accounts for 71.9%, while other factors account for 28.9%. This also indicates that the model is iterative, and the author hopes to have a better module to quantify these factors.

3.4 Limitations

This paper studies three models, namely multiple linear regression, random forest, and LSTM, to predict the future price of the same stock. However, there are some technical defects in code editing and model construction. Part of the code borrows from the existing model on Kaggle [9, 10]. The debugging of some model parameters may not be the best model data. In addition, owing to the length

of exposure to the field of mechanical learning, the impact of external factors are not investigated in depth, e.g., COVID-19 or the war in Russia and Ukraine. All model construction is based on the price-volume indicators of stocks.

4. Summary

In conclusion, this paper investigates stock price prediction based on OLS, Random Forest, LSTM. Stock price prediction has been a topic since the issuance of stocks. The iteration of the model has a positive impact on stock price prediction. This paper uses three models (i.e., multiple linear regression model, random forest regression model, and LSTM model) to predict the stock price of Johnson & Johnson. When the variables are only open, close, high, low price, multiple linear regression has better fitting effect. The three models have successfully predicted the stock price in the short term. However, ascribed to the time limit and researching depth in related fields, it is failed to establish a more advanced model with more variables. In the future, more professional analysis in related fields should be carried out in terms of different model regarding to the impact of COVID-19 on stock price. Overall, this study provides future researchers with a picture of the impact of COVID, a global epidemic, on stock forecasts. It is convenient for beginners of machine learning to understand and summarize random forest, RNN (LSTM) models.

References

- [1] Wang Peidong. Analysis and prediction of stock price based on multiple linear regression. Science and technology economic market, 2020, vol. 01, pp 90-91.
- [2] Kohara K, Ishikawa T, Fukuhara Y, et al. Stock price prediction using prior knowledge and neural networks. Intelligent Systems in Accounting, Finance & Management, 1997, 6(1): 11-22.
- [3] Yu P, Yan X. Stock price prediction based on deep neural networks. Neural Computing and Applications, 2020, 32(6): 1609-1628.
- [4] Hidalgo B, Goodman M. Multivariate or multivariable regression?. American journal of public health, 2013, 103(1): 39-40.
- [5] Shah D, Isah H, Zulkernine F. Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 2019, 7(2): 26.
- [6] IBM Cloud Education. Recurrent Neural Networks, IBM, Retrieved from: <https://www.ibm.com/cloud/learn/recurrent-neural-networks>
- [7] Erin Schumaker. Timeline: How coronavirus got started, abcNEWS, September 22, 2020, Retrieved from: <https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165>
- [8] Goh T S, Henry H, Albert A. Determinants and prediction of the stock market during COVID-19: Evidence from Indonesia. The Journal of Asian Finance, Economics and Business, 2021, 8(1): 1-6.
- [9] Pushkar Jain. Stock Price Prediction, Kaggle, Retrieved from: <https://www.kaggle.com/code/l33tc0d3r/stock-price-prediction/notebook>
- [10] Yash Patel. Advanced Stock Pred using SVR, RFR, KNN, LSTM, GRU, Kaggle, Retrieved from: <https://www.kaggle.com/code/ysthehurricane/advanced-stock-pred-using-svr-rfr-knn-lstm-gru/notebook>