Stock Daily Return Prediction of Amazon and Alibaba using Linear Regression and LSTM

Jinxian Lyu*

Department of Science, The University of Melbourne, VIC, Australia

*Corresponding author: jinxian@student.unimelb.edu.au

Abstract. The stock plays a key role in the economy market. As an individual investor, predicting stock return is always a hot research topic. In recent years, Amazon and Alibaba have become the head of e-commerce and the competition between them is becoming intense. In this paper, linear regression model and LSTM model based on machine learning are introduced and applied to predict the stock return of these two companies. RMSE and R^2 are used to choose the number of variables and evaluate those models. This study finds that the simplest linear regression is even better than LSTM model with limited sources. The negative R^2 scores of these two methods implies the nonlinearity and instability of stock return. However, predicting stock return is still possible. To investigate more about predicting stock return, more information such as the turnover rate and other non-numerical variables can be included in other models. Different types of LSTM model with different parameter setting can be applied to investigate deeply.

Keywords: Stock return prediction; linear regression; LSTM; amazon; Alibaba.

1. Introduction

The stock market is a significant part of every country. The flow of the stock market influences the virtual economy crucially. The stock price shows the expectation of the investors towards the market. Many researchers have done a lot of work to try to predict the stock price and stock market trend with diffferent dimensions [1-3]. Though some of those researches shows nice predictions of stock price, only using last day's stock price as the prediction also gives a nice result. This phenomenon is caused by the small price difference between two neighbour days.

As an individual investor, the favourite question is how much money will be earnd daily from each dollar invested. However, forecasting the daily return should be much harder than predicting stock price. The daily returns are always represented by percentages and depend on various factors, including numerical variables (close price, volume, turnover, etc.) and qualitative features such as policy and financial news. Besides, a certain company's stock is in a dynamic system that never varies alone. The movement of a certain stock can interact with other correlated stocks. Even if we predict the stock return successfully, the prediction itself will influence the stock owners' thoughts towards the stock. Predicting stock return remains to be one of the most demanding tasks in the machine learning field. Only the returns from previous days have been chosen as the basis of prediction as individual investors do not always have much information.

Amazon is now the largest e-commerce market in the US and is the second largest in the world, just lagging behind Alibaba [4]. Moreover, increasing amount of consumers favors e-commerce service as the electronic commerce grows rapidly [5]. Therefore, it's of great significance to make a prediction for Amazon's stock price. One of its most welcomed online sales services, Amazon Prime, now has more than 100 million members across the globe [6]. Alibaba is now the largest e-commerce platform in the whole world. Alibaba's Internet markets in China were home to 423 million active buyers with a total gross merchandise volume (GMV) of $485 billion [7]. Alibaba's 2018 GMV of USD 768 billion is more than triple that of Amazon, the world's second-largest e-commerce platform, at USD 239 billion [8].

Linear regression algorithm defines a linear relationship between the predictor variables and the response. The least square approach is widely used in estimating the coefficients. Dimensionality reduction along with cross-validation is implemented to avoid the overfitting problem. Recurrent neural networks (RNNs) are a subtype of neural networks which may update the current state based

on previous states and current inputs with the help of feedback connections [9]. The stock price forecast can benefit from the feedback mechanism of RNNs. With at least a single feedback connection, RNNs have the ability to learn correct patter from a time sequence. But the prediction is based on more recent history than the past, which can result in the problem of short-term memory [10]. Long short-term memory (LSTM) is a specific RNNs where prior-state information can be retained . LSTM was proposed in 1997, and this method introduced gates into the cell structure to deal with the problems of vanishing gradient and exploding gradient. During the training process, LSTM can learn the long-term dependency which is of great importance in the stock market.

In LSTM, The Forget door determines which information should be removed from the previous cell state. Input gate will update the input information. These two gates determine the current cell state. The output gate controls the number of outputs from the current cell state. Sigmoid function plays an important role in three gates. The value for the sigmoid function ranges from 0 to 1. The mechanism is that 0 means the information will be completely discarded and 1 means to be completely retained. There was a study showed that the linear regression is even better than LSTM when the data sources are limited [8]. This study is trying to see if linear regression and LSTM methods are useful on predicting the return and which is better. The different performances of return and price will be tried to explain.

2. Methodology

2.1 Data Structure

The data of close prices of Amazon from 1997 to 2020 and Alibaba from 2014 to 2022 were downloaded from the website called Kaggle. The files downloaded are in the format of csv (Comma-Separated Values). Both files contain the information of every day's open price, highest and lowest prices of the day, close price, and the date. As the close price can be seen as a summary of daily stock price, only columns of date and close price are elected to be investigated in this research. These two companies' close prices from November 2014 to July 2020 have been chosen to use because data in the same period are easier to compare. As the main keyword in this research is return, one more step which is converting close prices to return needs to be done. The daily return is defined as the difference between two days' close prices divided by today's close price.

2.2 Linear Regression

For linear regression method, both data sets are split into a training set, a valid set, and a test set in the ratio of 3:1:1. The plots of these data sets are shown in Figure 1 and Figure 2. The prediction results for the test set are subsequently used to measure the performance of the applied models on different data sets.

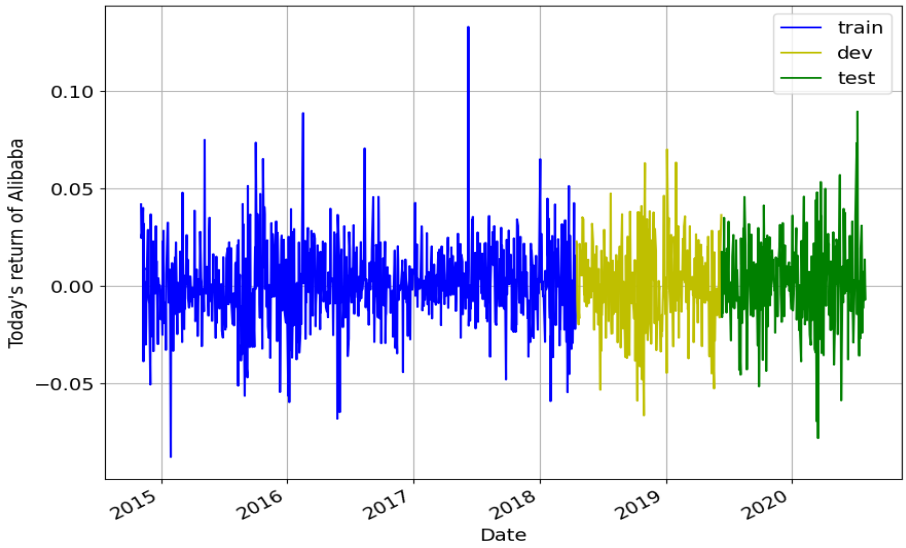

Figure 1. Daily returns of Alibaba from November 2014 to July 2020

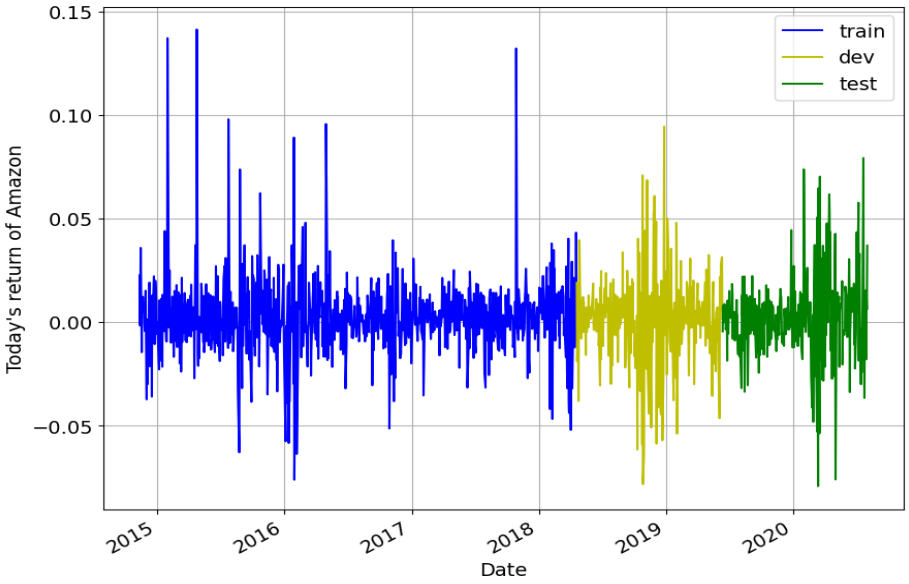

Figure 2. Daily returns of Amazon from November 2014 to July 2020

In Figure 3 and Figure 4, variable N represents the number of previous days used to form the linear regression model and R^2 represents the Coefficient of Determination. The evaluation of the valid set is used to decide how many previous days will be chosen to use as predict variables. In this study, the coefficient of determination and the RMSE (Root Mean Squared Error) are presented to assess the performance of predictions on datasets.

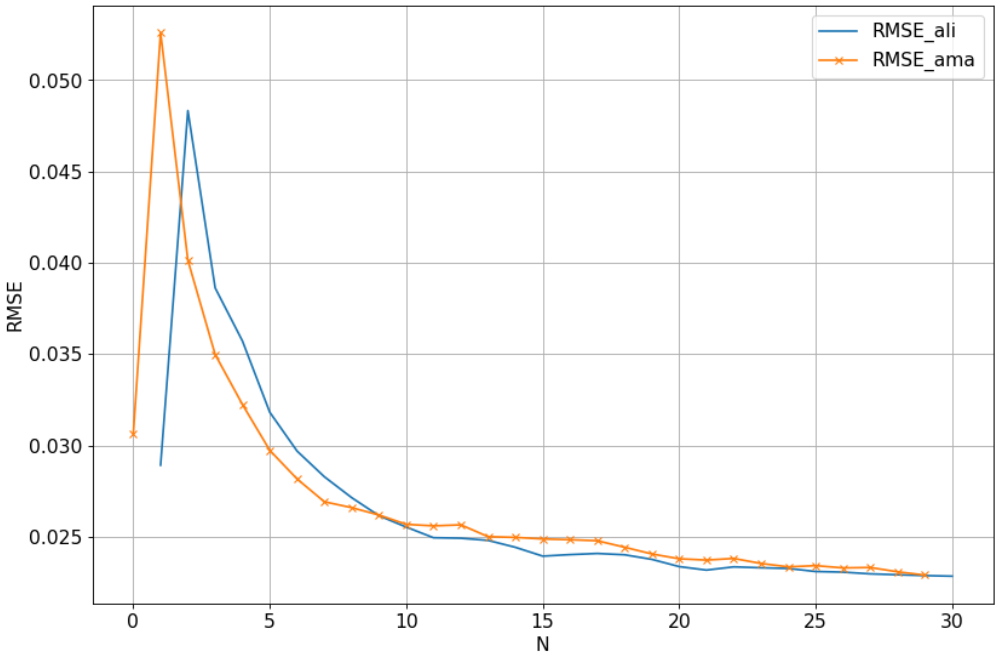

Figure 3. RMSE vs. number of days

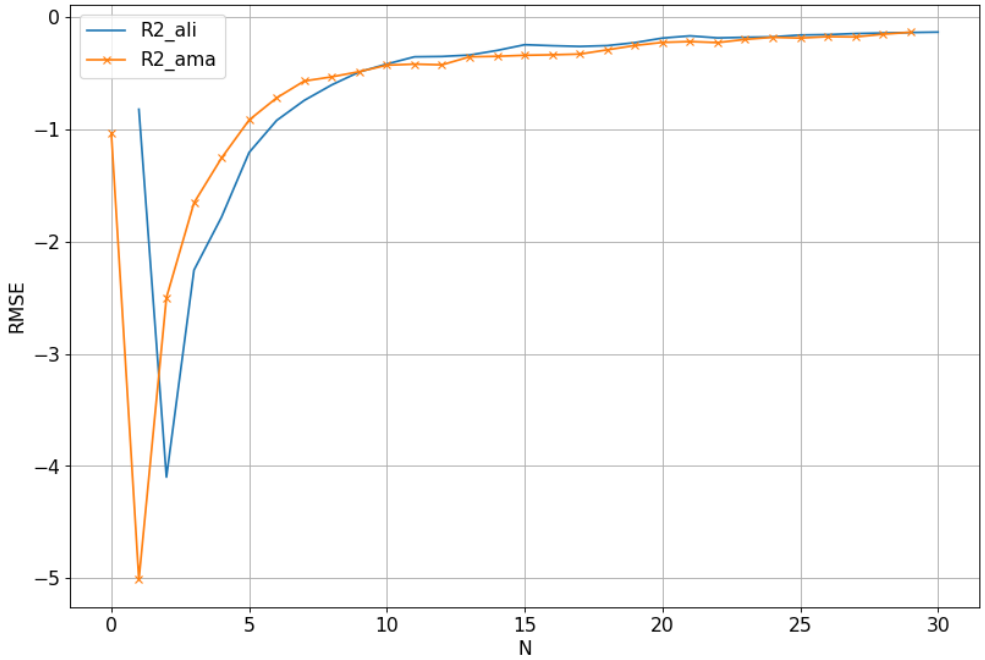

Figure 4. Coefficient of Determination vs. number of days

In Figure 3, RMSE decreases as N increases. RMSE of both companies starts to decrease slowly from N equals to 10. In Figure 4, Coefficient of Determination also increases with lower speed from 10. The goal is to maximize the Coefficient of Determination and minimize RMSE while avoiding the overfitting problem at same time. So, in this study's linear regression model, the close price of next day will be predicted through last 10 days' close price.

2.3 LSTM Model

For the LSTM model in this study, a sixty-dimensional array (as Table 1 shows) is used as input to predict the next day's return which is shown in Figure5. The first days are removed because returns are NaN.

Table 1. Daily return of Alibaba and Amazon used in LSTM

Date_Ali	Today	Date_Ama	Today
2014/11/4	0.042	2014/11/11	0.023
2014/11/5	0.025	2014/11/12	-0.002
2014/11/6	0.027	2014/11/13	0.016
2014/11/7	0.027	2014/11/14	0.036
2014/11/10	0.040	2014/11/17	-0.015
…	…	…	…
2020/7/27	0.007	2020/7/27	0.015
2020/7/28	-0.007	2020/7/28	-0.018
2020/7/29	0.014	2020/7/29	0.011
2020/7/30	0.001	2020/7/30	0.006
2020/7/31	-0.007	2020/7/31	0.037

2.4 Model Evaluation

The prediction performance is evaluated by the Root Mean Square Error (RMSE) and Coefficient of Determination (R^2-score). R2 is the proportion of the variation of the dependent variable which can be predicted from the independent variable [9].

$$R^2 = 1 - \frac{SSres}{SStot} = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (1)$$

The RMSE is a measure of accuracy, to compare the prediction errors of different models for a particular dataset and not between datasets, as it depends on the scale. [10]. The smaller the RMSE, the closer the predicted value is to the true value. However, it is not easy to interpret the difference caused by different models and types of data sets. Therefore, R^2-score can be a better indicator for model evaluation. R^2-score represents how the prediction fits the true data. It is better if R^2-score is close to 1.

3. Results and Discussion

3.1 Linear Regression

For linear regression model in this research, Y_i represents ith day's return and X_{ip} represents the daily return p days ago. As the number of days used for prediction is 10 in this paper, p is in the range of 1 to 10. The aim is to get the best estimates of β_p.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \;\; p = 1, \ldots, 10 \quad (2)$$

Figure 5 and Figure 7 shows whole test data with predictions for a year. To be more explicit, a proportion of result from each dataset is made into Figure 6 and Figure8. Figure 5 and Figure 6 show the green test data and the red predictions of Alibaba made by linear regression model. The RMSE of the test data of Alibaba is 0.027 and the R^2-score is -0.559. The RMSE of Amazon is 0.024 and the R^2-score is -0.457. The negativity of R^2-score means that the linear regression model does not fit the data given. These two RMSE will be compared with LSTM model later.

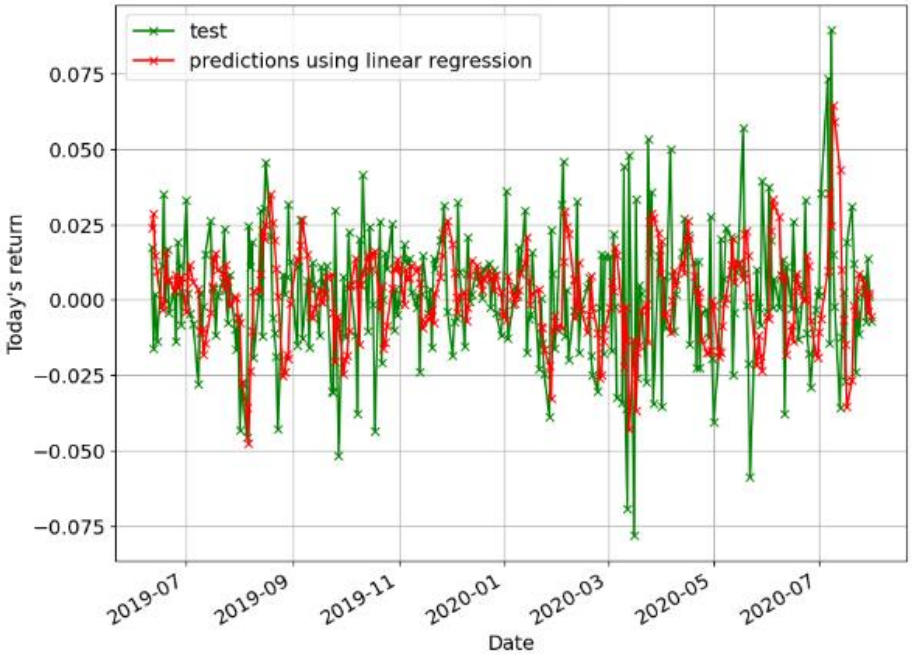

Figure 5. Result of Alibaba during 2019.6-2020.8

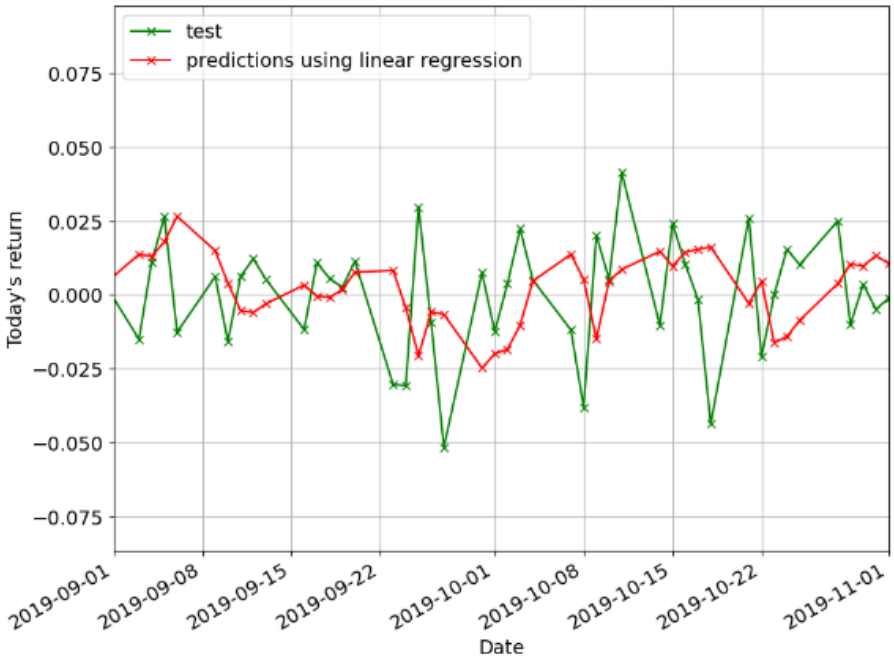

Figure 6. Result of Alibaba during 2019.9.1-2019.11.1

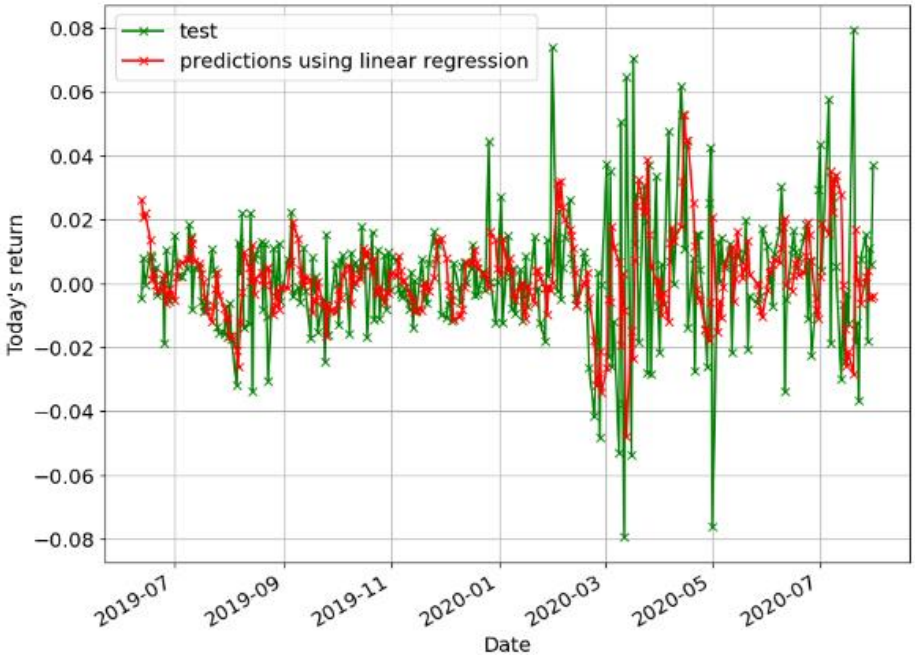

Figure 7. Result of Amazon during 2019.6-2020.8

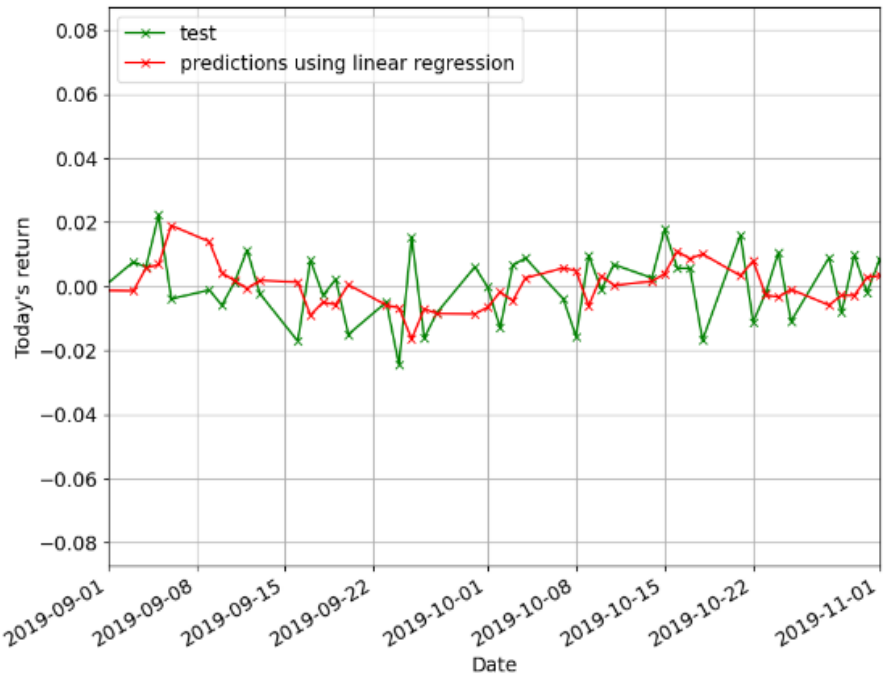

Figure 8. Result of Amazon during 2019.9.1-2019.11.1

Figure 7 and Figure 8 show the green test data and the red predictions of Amazon made by linear regression model. The RMSE of the test data of Alibaba is 0.027 and the R^2-score is -0.559. The RMSE of Amazon is 0.024 and the R^2-score is -0.457. The negativity of R^2-score means that the linear regression model does not fit the data given. These two RMSE will be compared with LSTM model later.

By paying attention to Figure 6 and Figure 8, which have shorter period compared to whole test data, some intuitive difference between predictions by linear regression model and the true values of daily return is apparent.

3.2 LSTM

Figures 9 and 10 include the predictions with the true values of return in plot and table forms are shown in Table 2 and 3.

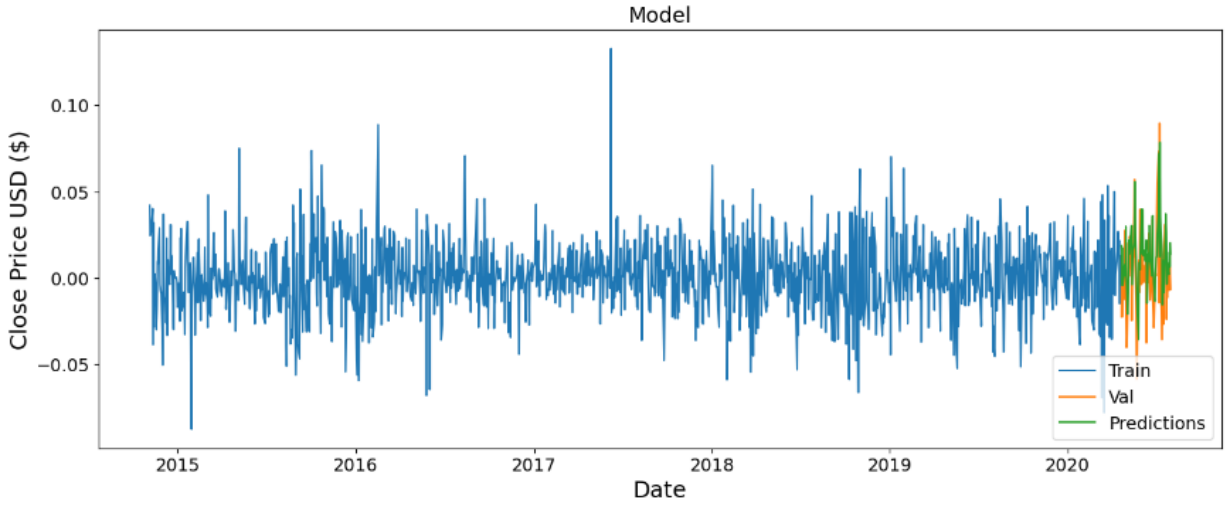

Figure 9. Result of Alibaba

Table 2. Result of Amazon (data set 2: 2018-2022)

Date	Today	Predictions
2020/4/21	-0.023	0.002
2020/4/22	0.013	0.004
2020/4/23	-0.022	0.017
2020/4/24	-0.004	-0.004
2020/4/27	-0.003	0.005
…	…	…
2020/7/27	0.007	0.002
2020/7/28	-0.007	0.015
2020/7/29	0.014	0.006
2020/7/30	0.001	0.020
2020/7/31	-0.007	0.014

From table 2, predictions and true values of daily return of the test set are labelled clearly. Even without evaluation indicators, the relative difference between prediction and actual return is obvious. By applying LSTM model to Alibaba data, the evaluating indicators are following: R^2 = -0.651, RMSE = 0.031.

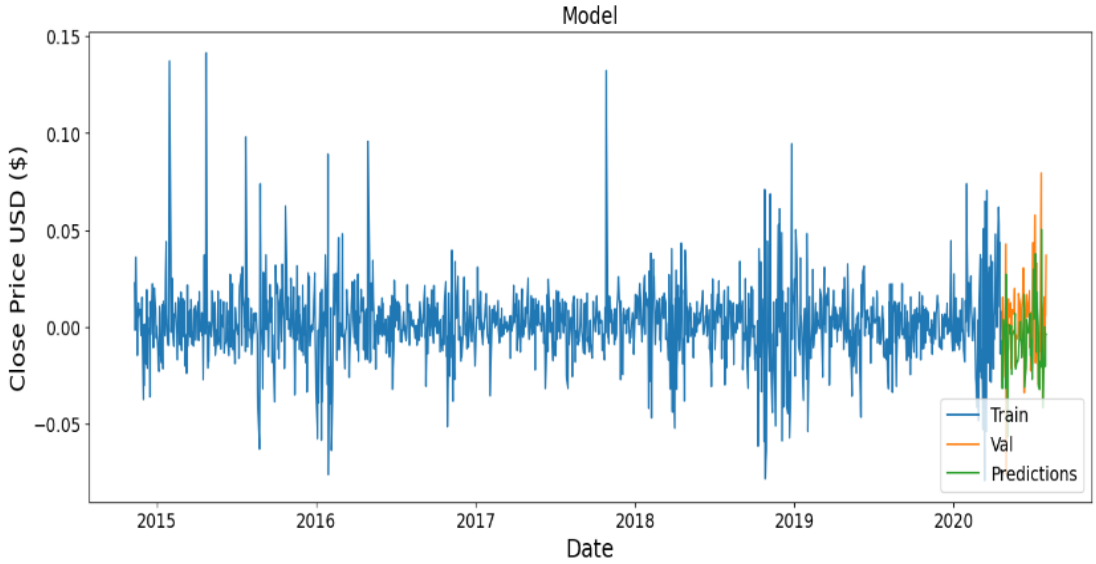

Figure 10. Result of Amazon

Table 3. Result of Amazon

Date	Today	Predictions
2020/4/21	-0.027	-0.007
2020/4/22	0.015	-0.032
2020/4/23	0.015	-0.005
2020/4/24	0.004	0.004
2020/4/27	-0.014	-0.001
…	…	…
2020/7/27	0.015	-0.015
2020/7/28	-0.018	0.000
2020/7/29	0.011	-0.020
2020/7/30	0.006	-0.004
2020/7/31	0.037	-0.004

By applying LSTM model to Amazon data, the evaluating indicators are following: R^2 = -1.031, RMSE = 0.032. Comparing RMSE and R^2 score of LSTM model with Linear regression model, Linear regression has the better performance.

4. Conclusion

This paper conducted an empirical study of linear regression model and LSTM on Alibaba's and Amazon's daily return. Overall, linear regression even works better than the LSTM model. However, both linear regression and LSTM model fit the data badly. It may be caused by the uncertainty and volatility of stock return. In addition, the non-linear feature of the stock market also limits the performance of linear regression models on stock price forecasting. The linear regression model assumes the stock return to increase or decrease linearly, but the actual values tend to jump up and down over the period. Also, various factors affecting stock prices are hard to quantify and cannot be included in the model.

For the directions of future work, more features need to be taken into consideration. In addition to numeric features, non-numeric factors such as policy and financial news are also important in the stock market. More Besides, various types of LSTM models can be chosen to see if they show the same results as this study.

References

[1] Manurung A H, Budiharto W, Prabowo H. Algorithm and modeling of stock prices forecasting based on long short-term memory (LSTM). ICIC Express Lett., 2018, 12 (12): 1277 - 1283.

[2] Irawan C, et al. Long Short-Term Memory Algorithm for Stock Price Prediction. 2022 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2022, 490 - 495.

[3] Wiiava A Y, Fatichah C, Saikhu A. Stock Price Prediction with Golden Cross and Death Cross on Technical Analysis Indicators Using Long Short Term Memory. 2022 5th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2022, 278 - 283.

[4] Adam Levy. The 7 Largest E-Commerce Companies in the World, The Mmtley Fool, 2018.

[5] Gaur P. Giants like amazon and alibaba are fueling the growth of e-commerce sector in india. PC quest, 2015.

[6] Shen S M, Wang H Y. Asymmetric Multifractal Analysis of the Chinese Energy Futures and Energy Stock Markets under the Impact of COVID-19. Fluctuation and Noise Letters, 2022.

[7] Bhattarai J K, Gautam R, Chettri K K. Stock Market Development and Economic Growth: Empirical Evidence from Nepal. Global Business Review, 2021.

[8] Zhang H. Stock Price Prediction using Linear Regression and LSTM Neural Network. 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), 2022, 182 - 185.

[9] Nittayagasetwat A, Buranasiri J. International Fund Flows and Anomalies in Asian Stock Markets. Asian Economic and Financial Review, 2022, 12.

[10] Bae J, Kang J. The negative hiring rate premium on stock returns in the Korean stock market. Pacific-Basin Finance Journal, 2022, 73.