August 31, 2020

Advanced Stock Price Prediction using LSTM Network for Royal Dutch Shell

UHSPE Machine Learning Bootcamp 2020

Team 1: Kalu Divine - Nhung Nguyen - Jiajun Chang - Jose Alvarado

Mentor: Pushpesh Sharma

1. Introduction

With the development of the computer parallel computing ability, more and more deep-learning algorithms and structures which includes vast amount of parallel arithmetic are applied to financial and other industries. In previous project, linear regression techniques are used which requires linear relations between input and output. While for stock price prediction, the price is determined by many factors many of which are implicit and non-linear. Thus, deep-learning structures are used to solve this problem. This project is an advanced version of the previous project.

In this project, the dataset used in this project and the data-processing method are introduced first. The Long Short-Term Memory Model, which is a type of Recurrent Neural Network (RNN), is introduced and the configuration of the model is elaborated. Finally, the result of the prediction is presented and the accuracy of the model is analyzed. The comparison between different models are drawn and the conclusion will be shown.

2. Dataset

These most representative data are mostly open-source online which can be accessed by the following methods:

- Through Kaggle datasets search engine(https://www.kaggle.com/datasets)
- Using datareader in pandas package in Python to fetch the data from websites
- Through other websites.

Based on this data, numpy arrays can be converted from this structure and Neural Networks can be based on the arrays.

3. Feature and Processing

The dataset used in this project has a time duration from 1987 to 2020. In previous project, only 2 recent years record are used which turns to mediocre prediction accuracy. Thus, in this project, the 10 years record will be used. The raw dataset will be stored into excel file which can be read through Python function and store as numpy structure dataFrame. After pre-processing, the head of the dataFrame is shown as follows:

	date	open	high	low	close	volume
0	2009-01-02	52.05	53.74	52.00	53.53	296100
1	2009-01-05	51.80	53.91	51.80	53.33	756000
2	2009-01-06	53.96	54.77	53.00	54.18	826200
3	2009-01-07	53.33	53.66	52.26	53.00	502800
4	2009-01-08	53.48	54.36	52.71	54.26	309400

Figure 1. Data frame structure that stores the stock price details

The close price of the stock will be chosen to represent the daily price. The time-series of the price is plotted as follows.

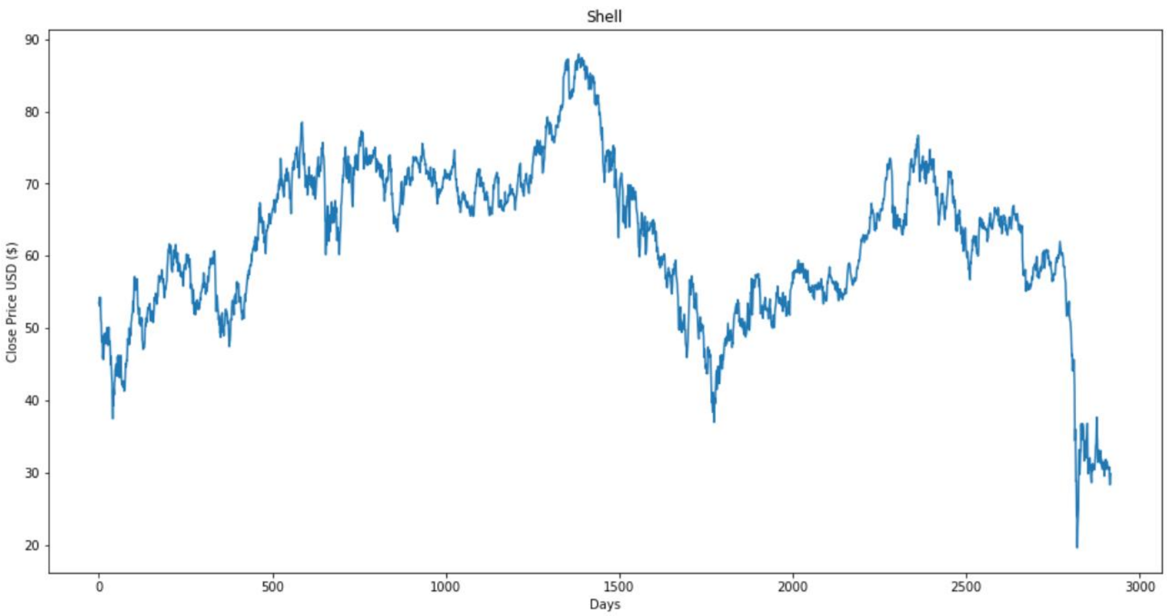


Figure 2. Time-series plot of the stock close price of 10 years period

4. Model and Techniques

When dealing with time dependent dataset, Recurrent Neutral Networks (RNN) is a better fitted technique to use. However, basic RNN remembers things for just small durations of time. When the variable in a dataset depends on a longer history values, the accuracy of the model decreases. This issue can be resolved by applying an advanced version of RNNs – the Long Short-Term Memory (LSTM) networks.

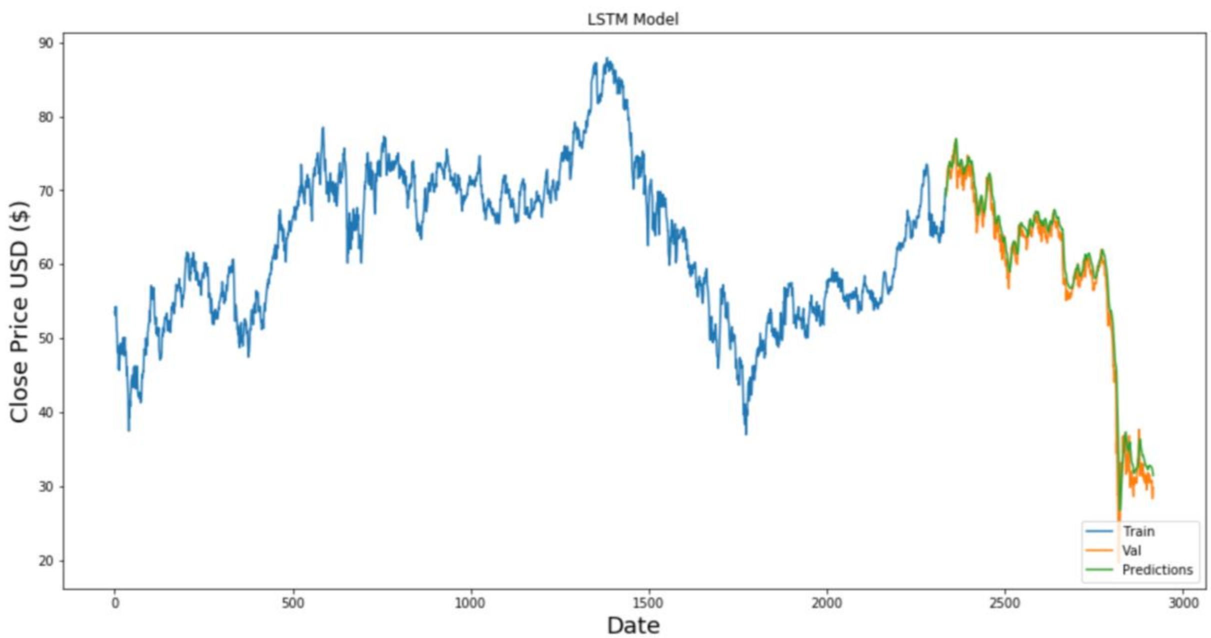
LSTMs makes small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things.[2] LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. It trains the model by using back-propagation.[1]

A LSTM network expects the input to be in the form [samples, time steps, features] where samples is the number of data points we have, time steps is the number of time dependent steps that are there in a single data point, features refers to the number of variables we have for the corresponding true values.[2] We also need to rescale the input values between 0 to 1. A sequential model which is a linear stack of layers is used. The first layer is an LSTM layer with 2247 memory units and it returns sequences. This is done to ensure that the next LSTM layer receives sequences and not just randomly scattered data.[2]

5. Results and Discussion

The smaller the mean square error, the more accurate the prediction values. Predicted values estimated by LSTM Model mean square error of 1.38 when predicting the last quarter of the dataset. The predicted trend reflects the actual trend of the dataset.

This result is much better than expected. Comparing to previous models used in project 2, the confidence level of predicted prices by LSTM model is higher. Using the same process, the model is observed to predict more accurate values when it is trained with a longer history dataset.



LSTM RMSE: 1.38187559

Figure 3. Predicted prices using LSTM model and the mean square error

6. Conclusion

LSTMs are a very promising solution to sequence and time series related problems. The code of this project can be developed to be an efficient feature for a stock trading website. Additional modification will allowed this program the ability to give users a quick quote for their interested companies' stock price.

7. References

[1] Aditi Mittal (2019). Understanding RNN and LSTM. Retrieved from https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e

[2] Pranjal Srivastava. (2017). Essentials of Deep Learning : Introduction to Long Short Term Memory. Retrieved from https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/

[2] Royal Dutch Shell - 33 Year Stock Price History. Retrieved from https://www.macrotrends.net/stocks/charts/RDS.B/royal-dutch-shell/stock-price-history

[3] Stock Price Prediction Using Python & Machine Learning. Retrieved from https://www.youtube.com/watch?v=QIUxPv5PJOY&t=798s