Title: Stock Price Prediction using Long Short Time memory network

Group Members: Anlan Chen, Yunfei Fan

Group Name: Spicy Sugars

Abstract

In this report we are trying to use Long Short Term Memory (LSTM) network to solve stock price predicting problem. Historical stock price of 4 years of selected company is used to train and test LSTM model. We set the model with 4 LSTM blocks and 1 output payers, with 96 units in each block. Time step of 20, 40, 60, and 80 is examined to select one with minimum mean squire error (MSE) using cross-validation strategy Time Series Split (TSS). The result indicate time step of 60 minimize MSE with a value of 0.4225. According to the visualized plot of predicted and original stock price, our trained model did a fine job on predict close price with similar trend of increasing/decreasing, although a small delay can be observed

Problem description

Our problem is to predict stock price of semiconductor companies using machine learning algorithms. We will use historical data on the stock price of a publicly traded semiconductor company to predict the future stock price of this company. We hypothesize the historical data could reflect the future trends in some extent, but we are unable to hypothesize whether there is an increase or decrease trend without analyzing specific historical data.

This is an interesting question because market forecasts could offer huge profits. To forecast the market, most researchers use technical or fundamental analysis. Technical analysis focuses on analyzing price direction to predict future prices, while fundamental analysis relies on analyzing unstructured textual information like financial news and statements. If we can use technical analysis methods, such as machine learning, to predict future stock price movements, then we can use some financial products to make buy and sell decisions and gain huge benefits, which is

also the potential impact of solving this problem. Other people will also be interested in this problem, since many people want to know information of how market will go and get a higher return of investment, especially those investors without professional knowledge.

There has been lots of researchers and analysts predicting stock price using machine learning in financial industry. Researchers in Analytics Vidhya used k-Nearest Regression and Long Short Term Memory to learn historical data of Tata Global Beverages. [1] Faculties in Umia University also used several methods to do stock price prediction, and they found deep learning is better in prediction, and the support vector regression method is the second best approach. [2]

Data description

We are using historical Nasbaq Real Time Price of stock 'QUALCOMM Incorporated (QCOM)' from April 2018 to April 2022 as training and validation data. The data comes from Yahoo Finance online, which open to public and can easily get access to. The cvs file is directly downloaded and is ready to be read. The raw cvs file include seven columns of date, open, high, low, close, adj close, as well as the volume information. In our research, we predict 'Close' price, which is adjusted for splits, instead of 'Adj.Close' adjusted for splits and dividend and capital gain distribution. In the Long Short-Term Memory(LSTM) model we tried, we are using historical close data to predict future close data.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2018-04-03	54.290001	55.049999	53.619999	54.779999	48.340557	7930700
1	2018-04-04	53.799999	55.119999	53.419998	54.990002	48.525875	7496700
2	2018-04-05	55.490002	55.500000	54.439999	55.040001	48.569996	5823300
3	2018-04-06	54.419998	54.770000	53.110001	53.119999	46.875687	8322200
4	2018-04-09	53.520000	54.889999	53.340000	53.430000	47.149254	7608500

Figure 1. raw data

Preprocessing

A simple visualization of extracted 'close' value as a function of date is shown in figure 2.

We first sort the data in ascending order and then create a separate dataset to prevent any new feature affecting the original data.

LSTM model: To fit to LSTM model's need, only 'close' column is extracted as df. According to our review of cases on the internet, Minmax scalar is the most frequently used scalar for similar problems, so we also applied Minmax scalar in our project. In the second half of our work, we add the cross-validation base on the previously mentioned model. Rather than using the 'regular' cross-validation method to splitting the data, we adopt the Time Series Split (TSS) function to get rid of the correlation. The detail of data splitting is discussed in method section.

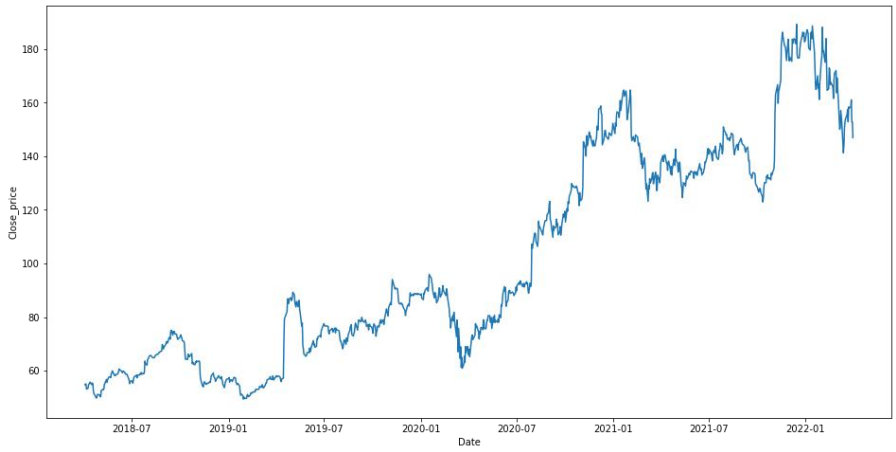

Figure 2 raw close price data

Method

The problem we are facing is a regression/forecasting issue. Long short-term memory module is a recurrent neural network (RNN) design to solve sequence dependence time series prediction problem. Different from general RNN, LSTM have multiple neural network layer and gates, as shown in figure.3. This enable it to study the data pattern for a long period of time (longer memory), so we decide it could be

a good candidate for the stock price predicting problem. We picked this module also for its simplicity of only using close price, and ability to process entire sequence of data using feedback connections.

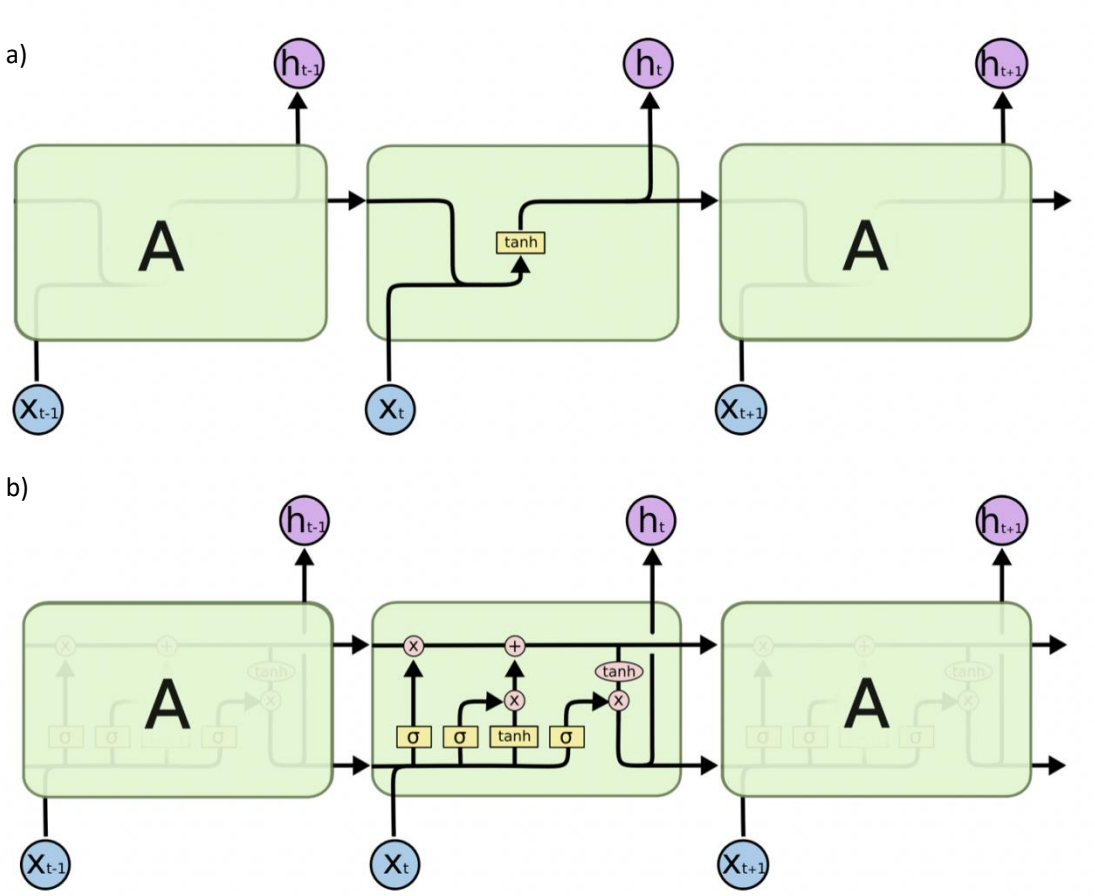

Figure 3 a).The repeating module in a standard RNN contains a single layer. b). The repeating module in an LSTM contains four interacting layers.

In the progressing report, we built a primary LSTM model without cross-validation. To create X and Y sets from previous train and test dataset, a 'look back' time step i is defined to determine numbers of X values (historical 'close' price) to be analysis for each Y (future 'close' price). In our project we tried i of 10, 30, and 60, and picked 60 according to its smallest validation MSE. Next comes the he core task of model building. We have built our model with 4 LSTM clocks and 1 output layer, with 96 units in each block. A dropout fraction of 0.2 is set for each block.

In the second half of our work, We also examine the optimum 'look back' (timestep) by circulate over nlag of 20, 40, 60, and 60 steps. Due to the joining of TSS

and timestep cycle, the model building cell is much more complicate than that in progressing report. Timestep cycle is set as outer cycle, in which df is split into x and y according to nlag value. TSS split is cycle is set as inner cycle, here x and y are split into training and testing set. At the end a dataframe is construct to select number of lags with minimun MSE.

Because we are dealing with regression problem, model performance is tested by MSE, 5 epochs are fitted. This is a smaller number than 50 we used before because otherwise running time will be unnecessarily long with cross-validation.

Result

Model with different nlag values are trained on training set, validated on cross-validation, and saved. According to the dataframe, timestep=60 corresponding to the minimum MSE with a value of 0.004225.

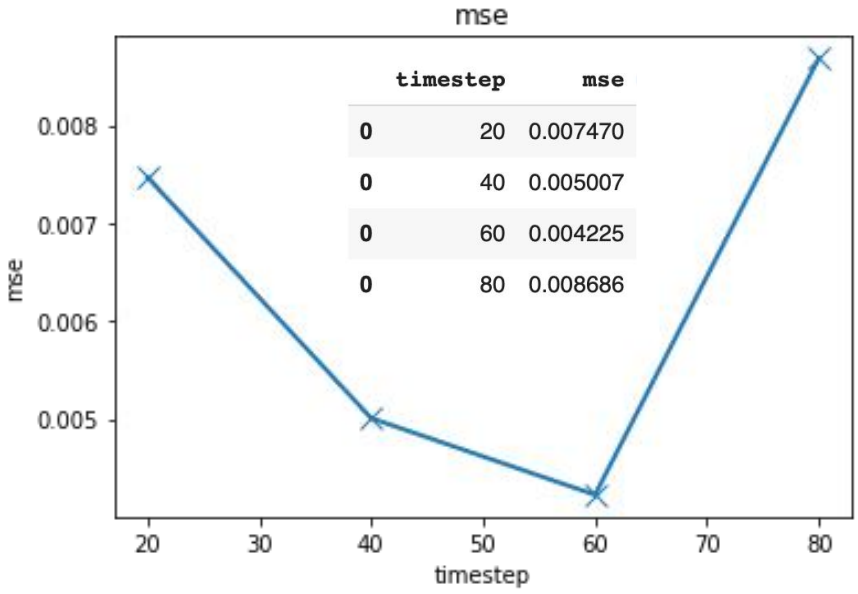

Figure 4 MSE and corresponding lags

The selected model is then applied on a freshly split test dataset and visualized, as shown in fig.4. The model did a fine job on predicting, trend of increasing/decreasing is basically correct, but some delay can be seen.

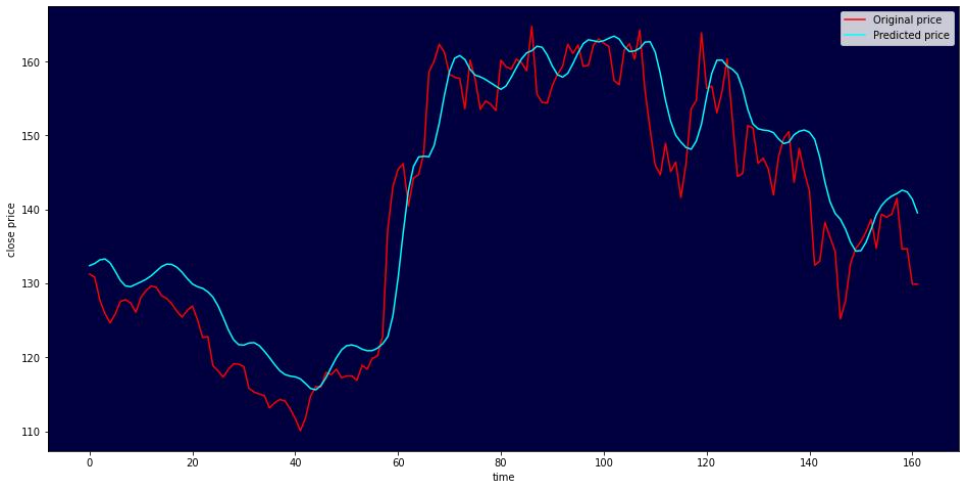

Figure 5 Visualization of predicted close price and original close price

Conclusion

It is important to predict stock price to make investment and stock buy and sell decisions. To this end, our LSTM model with four blocks, 96 units in each block, and one output layer works well in predicting close price, therefore, investors, managers, and market analyzers are recommended to use this model. When the predicted price starts to go up, buying is favorable; when it begins to go down, selling is recommended. Also, The result of this project could help investment companies and other users to make decision on portfolio in evaluation of the profitability in selling and buying stock.

Although the predicted close price in our LSTM model is roughly similar to the actual price, an unexpected shortcoming is obviously that we cannot know the trend of close price far in advance. Further analysis could focus on increasing the predictability of our model by manipulating the number of layers, units, and epochs.

Finally, we do not model and analyze the strategy return in this project. In the future, it is recommended to use log-transformation of time series differences to compare daily gain or loss of the specific strategy with that of S&P 500, and make decision on strategy selection.

Broader Impacts

Our LSTM model does capture some characteristics of stock price trend, but it is certainly not enough to identify whether the stock price will increase or decrease just by using LSTM model. Stock price is also affected by the recent news about company and other events like political events, or news of competitors. Therefore, it is difficult to accurately forecast stock price trend, and there will be some unethical institutions using the results of the model to cheat investors out of their money. It is important to identify the correct impact of machine learning to society .

Honor code

We adhered to the honor code in the completion of the assignment.

Reference:

[1] https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/

[2] https://onlinelibrary.wiley.com/doi/epdf/10.1002/isaf.1459

[3] https://data-flair.training/blogs/stock-price-prediction-machine-learning-project-in-python/

[4] https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[5] https://en.wikipedia.org/wiki/Long_short-term_memory

[6] https://stats.stackexchange.com/questions/241985/understanding-lstm-units-vs-cells

[7] http://colah.github.io/posts/2015-08-Understanding-LSTMs/