ISE 537 Final Project

Project 2: Price Prediction:

Comparison between ARIMA Model and LSTM

Boyue Wang 2090953408

Introduction of the project:

For this project, I decided to choose two stock which are in the same industry. The first two companies to show up in my mind are Pepsi and Coca-Cola. Both of them are head company in the beverage industry with profound corporate culture and history. Moreover, these two companies are competitors and have similar products. After having a basic glance of the two stocks, I decided to use these two stock for my price prediction project.

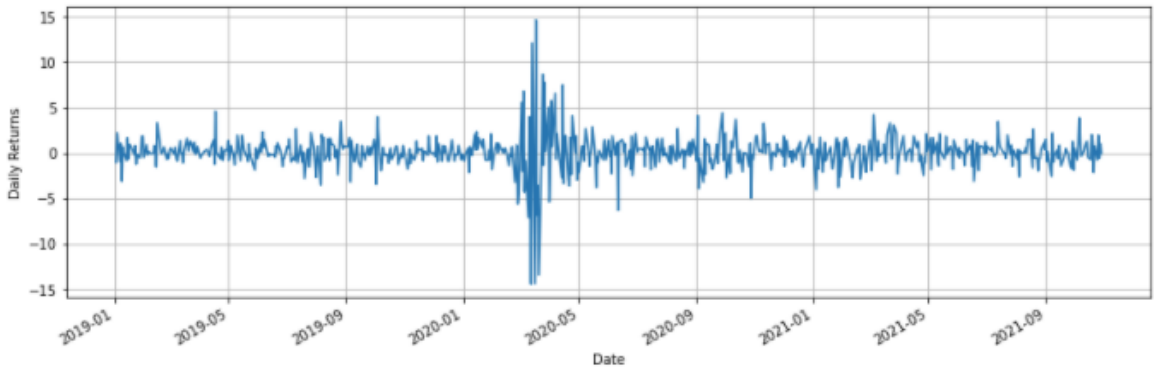
The time period I decided to pick in the end is from Jan 1st, 2019 to Oct 30th 2021. There are two main reasons that I select this period. First, since the project is stock price prediction, if the time period is too long, the prediction of the stock price would be meaningless. The financial market will have significant changes. In order to ensure the sample size is sufficient, and the time period is not too long. I used less than 3 years to do this project. Second, the COVID-19 had a significant influence in the stock market. I am curious that whether the prediction model can survive under the influence of COVID-19. In this way, I choose the time period which the COVID-19's time period is in the middle of my whole dataset. The dataset constructed by one year period before the pandemic, 3 months of the significant time period and one and half year period after the pandemic.

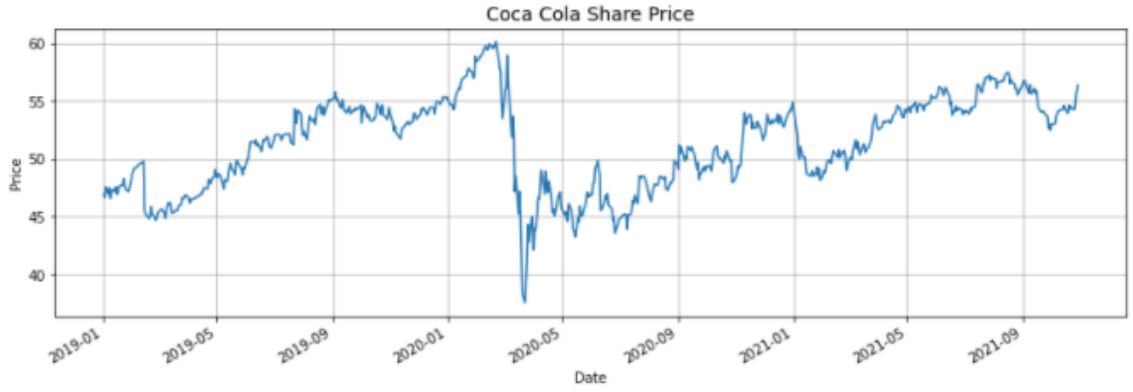
After decide the stock and time period, I import the data from Yahoo Finance. I will only use the close price of the end of each day through my time period. The time period I use is from Jan 1st, 2019 to Oct 30th 2021. In this way, each of the stock dataset have 714 data. I used 80% of the data to train the model and the rest of 20% of the data to test whether the model is predictable or not.

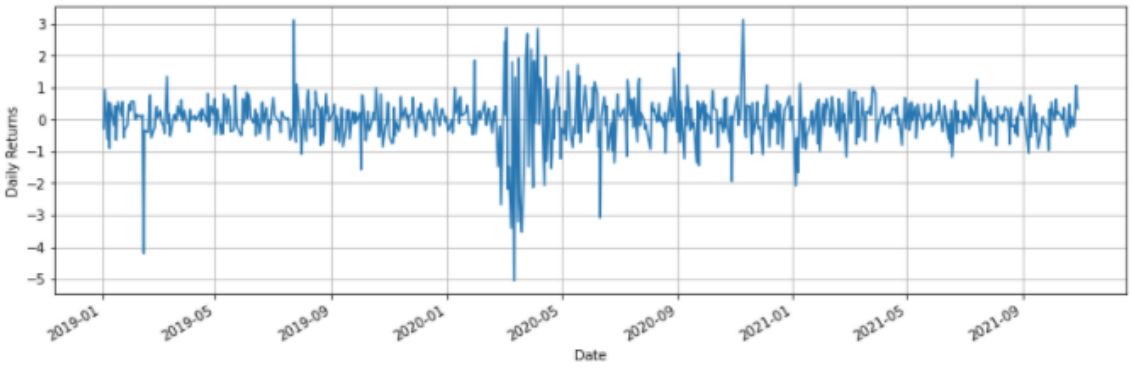
ARIMA Model:

ARIMA, full name Auto regressive Integrated Moving Average, is a class of models that use its own past values to predict future values of time series. The values are the lags and the lagged forecast errors of a given time series. However, there are too many limitations for ARIMA model. Only non-seasonal which are station time series can be modeled by ARIMA. ARIMA model is a combination of Moving Average(MA) model and Auto Regressive(AR) model. An AR model is purely depending on it own lags. A MA model is purely depending on the lagged forecast error.

Since the ARIMA model can only predict stationary time series. The first thing we should do is to make both Pepsi stock time series and Coca-Cola stock time series stationary. After preliminary test of the both stock. I find that the original Pepsi stock and Coca-Cola stock is not stationary. Under the test using adfuller, the p-value of Pepsi stock is 0.3 and the p-value of Coca-Cola is 0.13. Both of them are larger than 0.05. In this way, I decide to use the daily return of the to stock. After taking the difference of the two stock. Both of them passed the adfuller test and the p-value for both of them close to 0. Here is the plot for the stock price and the daily return of the two stocks.

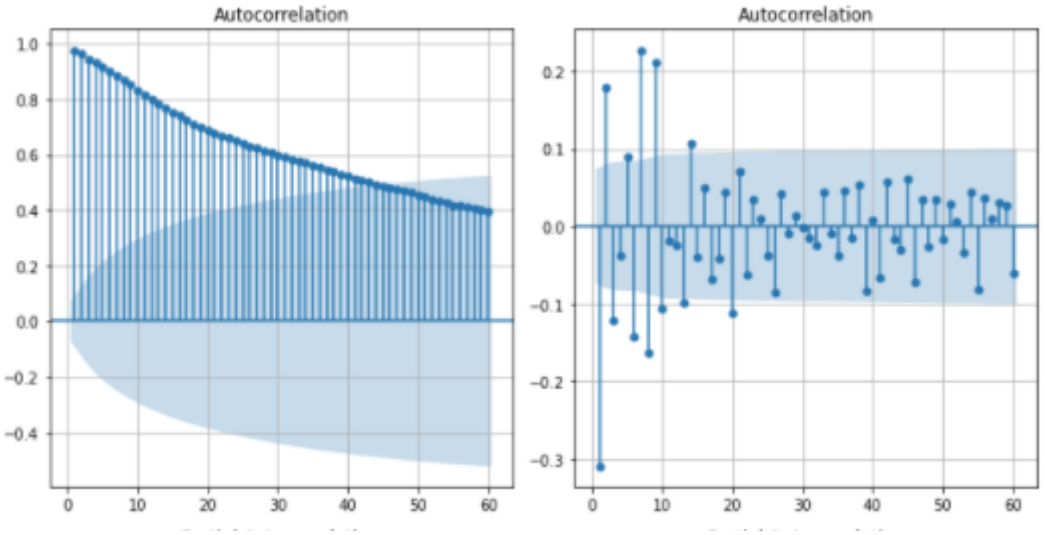




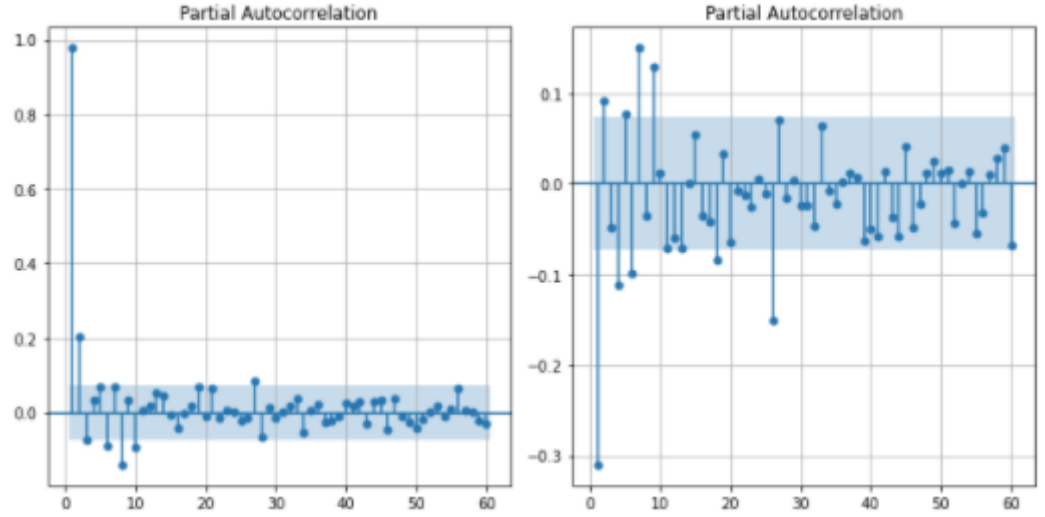


After making the two stock stationary, it's time to decide the p and q values. As we know, the partial autocorrelation plot describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information and the Partial autocorrelation plot only describes the direct relationship between an observation and its lag. We can tell whether we need to add MA terms from the autocorrelation plot. From the partial autocorrelation plot, we know we need to add AR terms. Here is the ACF and PACF for both stocks.

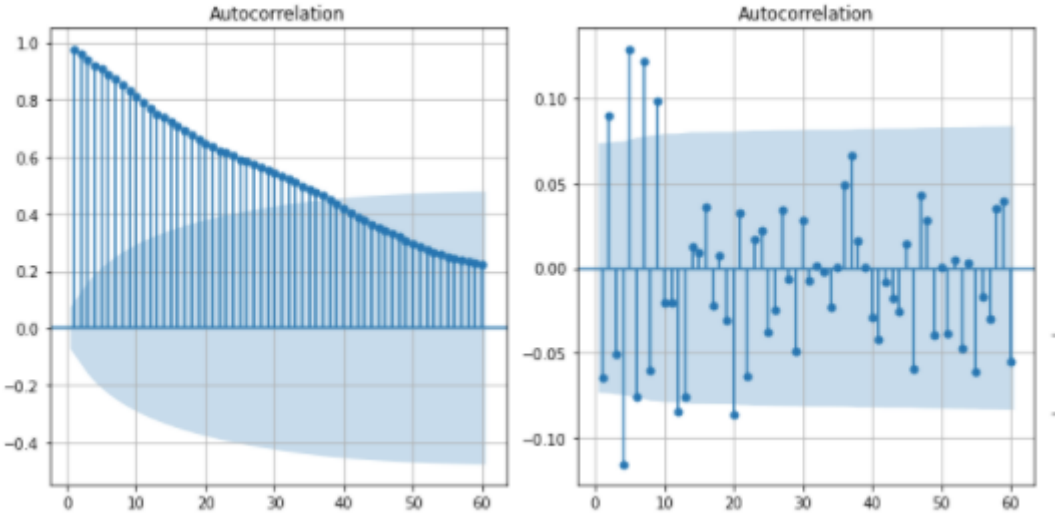
Autocorrelation for Pepsi



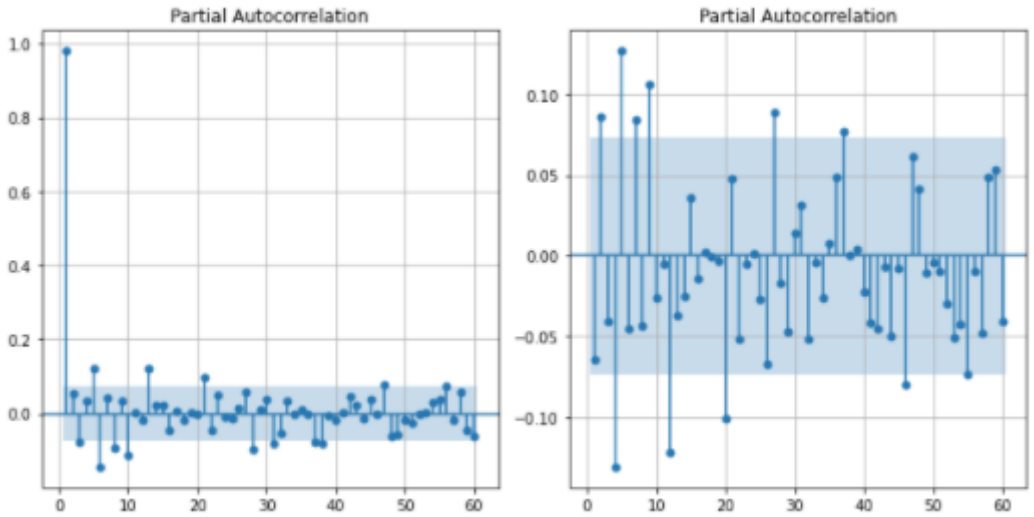
Partial Autocorrelation for Pepsi



Autocorrelation for Coca-Cola



Partial Autocorrelation for Coca-Cola



For both stocks, the lags before 10 are quite significant. However, we can not decide p and q terms only through the plots. For more precise calculation, I used searchARMA function from the class to find the smallest AIC and BIC using different p and q.

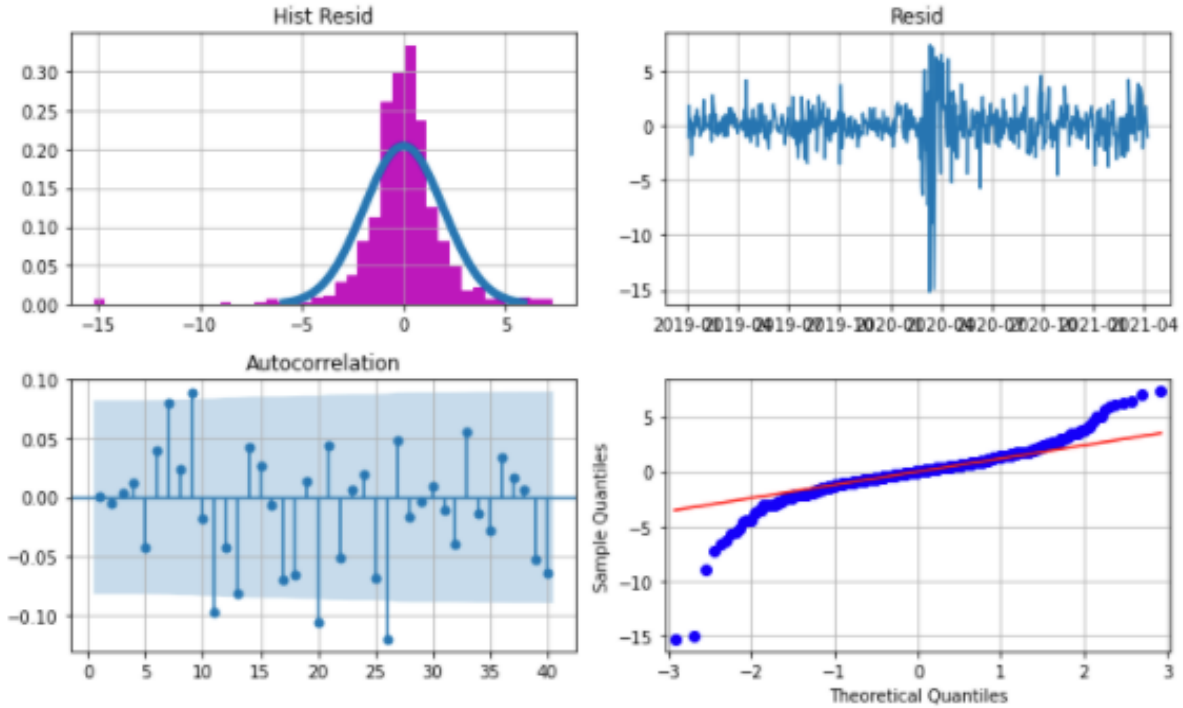
Pepsi:

I predicted the Pepsi stock price first. Using the searchARMA function, I found that for Pepsi stock, there is not a minimum value for both AIC and BIC. However, I select some p and q combination which has quite low AIC and BIC which showed below. To further decided the best p and q combination, I decided to test the root-mean-square error for each combination. Here is the chart of the result I have.

[p,q]	AIC	BIC	RMSE
[3,1]	2902.53	2929.95	5.94
[3,3]	2888.21	2924.77	6.03
[4,1]	2892.37	2924.36	5.91
[4,4]	2887.10	2932.80	6.05
[4,5]	2885.69	2935.95	5.97
[4,6]	2885.60	2940.43	5.96

From the chart above, we can find that the difference between the root-mean-square error is quite small. However, the combination [p,q] that I decided to choose is [3,1] since it has the smallest RMSE.

After running the model, and have a look at the residuals, we found that the mean of residuals is around 0 and there is no significant lag in the Autocorrelation plot. Also, the residual seems normally distributed. Although, from the Q-Q plot, we find the residual did not align on the line, it described the aggregated normality of the finance market.



From the plot below, we can see the out of sample prediction of Pepsi stock price. The actual Pepsi stock price is between the 95% confidence interval, which is good. Also, the prediction show the trend of Pepsi stock price increase.

Coca-Cola:

Then, I build the ARIMA model for Coca-Cola stock. To find the perfect p and q, I did the same thing as I found the p and q for Pepsi stock. I calculated the AIC and BIC for p and q in range 8, and find the root-mean-square-error for the selected p and q in the previous step. Here is the chart of p and q values for Coca-Cola stock.

[p,q]	AIC	BIC	RMSE
[1,4]	1576.22	1608.21	1.723

[2,4]	1577.59	1614.14	1.718
[3,6]	1571.39	1621.66	1.867
[4,1]	1576.21	1608.19	1.741
[4,5]	1571.23	1621.49	1.900

From the chart above, we can find that when [p,q] is [2,4], the root-mean-square-error has the smallest value. In this way, I choose [p,q] to be [2,4] and build the ARIMA model for Coca-Cola stock.

After running the model, we have the same result from Pepsi stock, the mean of residuals is around 0 and the distribution plot looks the same. Also, there is no significant lag fro autocorrelation plot.

After plot the out-of-sample prediction of Coca-Cola stock price, we find that the prediction predict the trend of Coca-Cola stock price. The stock price increase rate is not high as Pepsi stock price. Moreover, the actual Coca-Cola price is between the 95% confidence interval of the prediction of stock price.

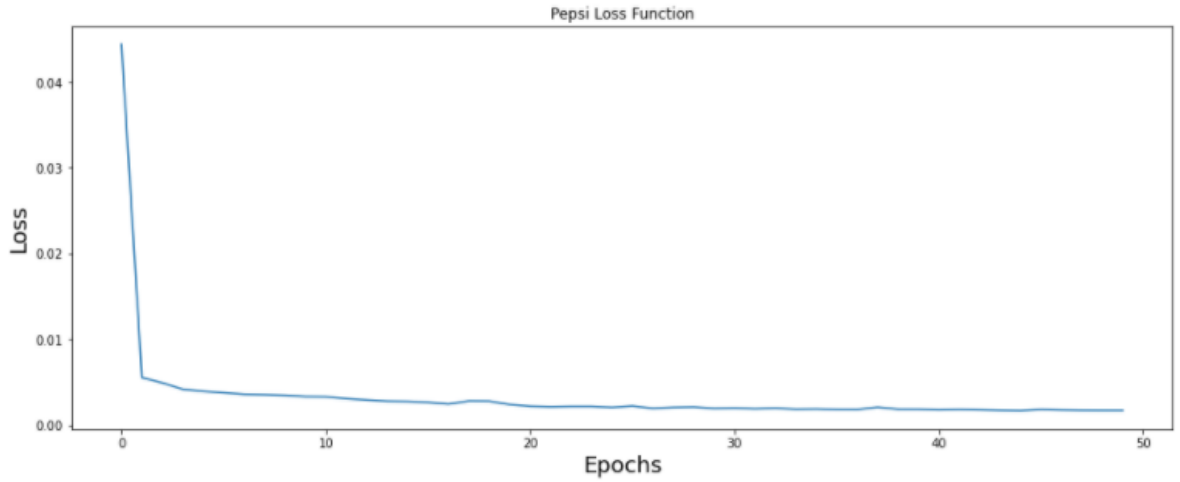
LSTM Model:

LSTM, short for Long short-term memory, is an artificial recurrent neural network(RNN) architecture used in the field of deep learning. Different from regular feedforward neural networks, LSTM has a memory capacity and are able to store data over a period of time, which can process arbitrary sequences of inputs. This characteristic make LSTM model extremely useful when dealing with financial time-series data. When we are using the LSTM model, we can have a choice of keeping the useful information and remove the others.

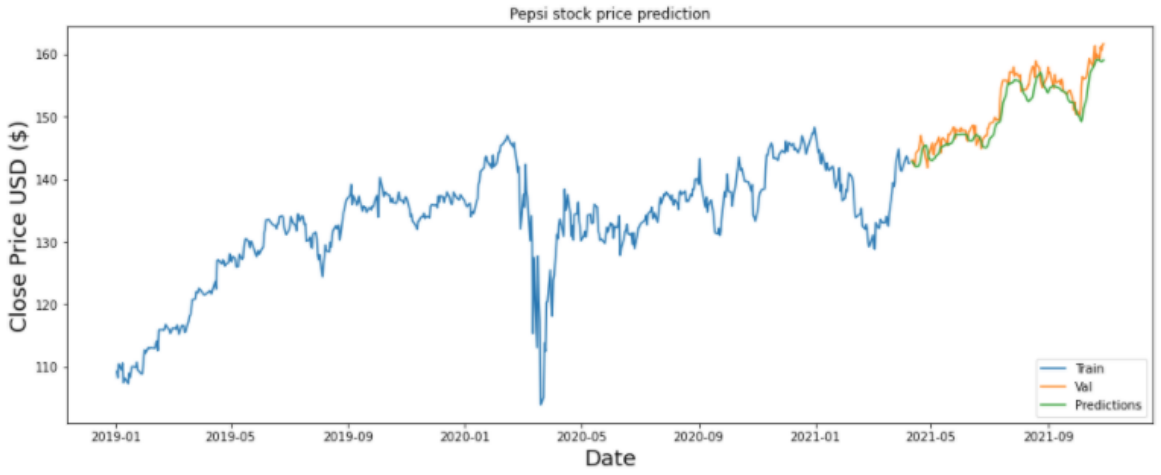
Same as ARIMA model, I still decided to use 80% of the data to test the model, and the rest 20% to test the model. Then I will build the LSTM model for Pepsi stock. The model

parameter is the same as the class LSTM example. Build the LSTM model with 2 hidden layers, the first one with 128 neurons and the second one with 64 neurons. Then assign 25 neuron in the first output layer and 1 neuron in the second output layer for predicting the normalized stock price. Also, the model will use MSE loss function and the Adam stochastic gradient descent optimizer.

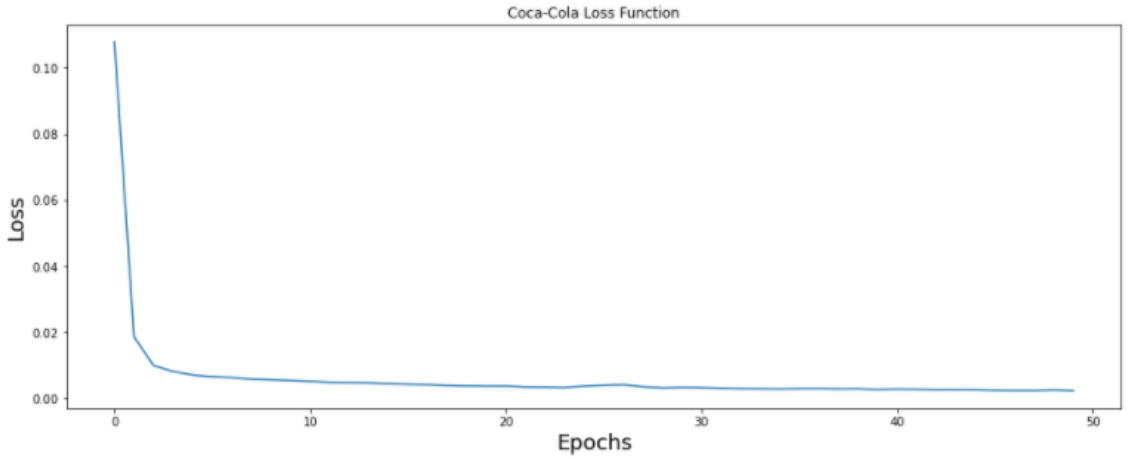
After we train the model with the Pepsi train dataset, we can produce the loss function for the train model below. From the graph below, we can find that the loss function is converging. In this way, we can say that the Pepsi LSTM model is convergence.



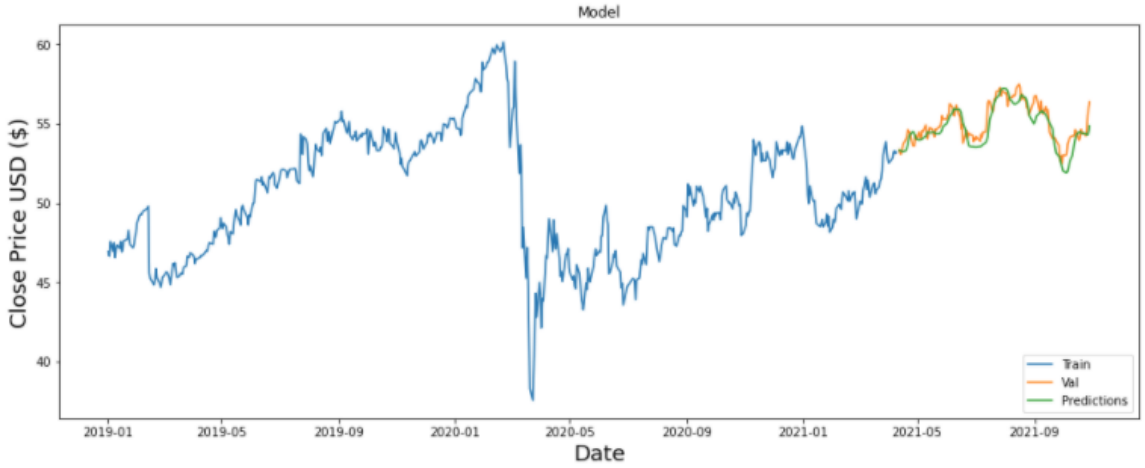
Moreover, the train LSTM model root-mean-square-error is 3.195 and the test LSTM model root-mean-square-error is 4.475. Although the RMSE of test model is larger than the RMSE of train model, the predicted result is acceptable since they are on the same level. The prediction of the test data is shown below, which is quite a good prediction.



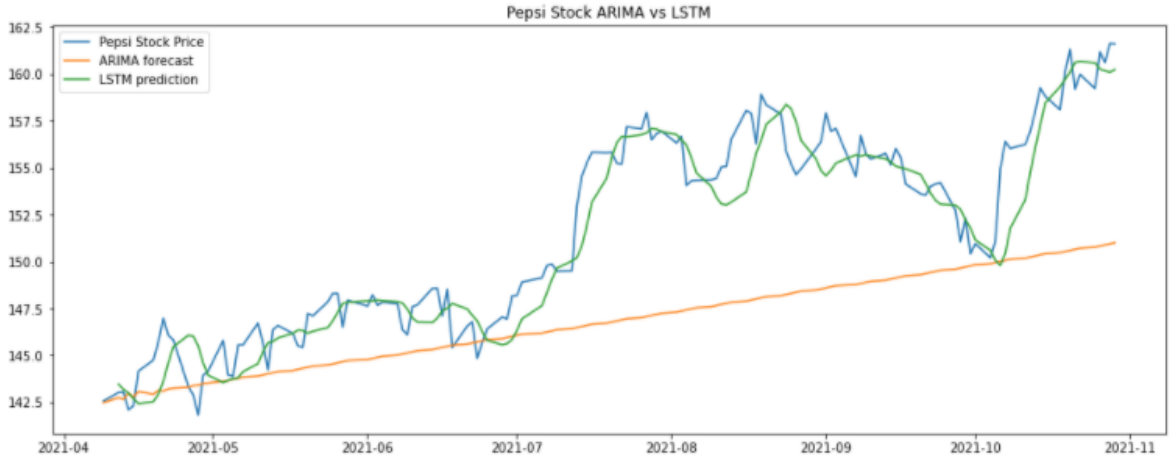
Then, I did the same process for Coca-Cola, and get the similar result. The train LSTM model root-mean-square-error is 1.07, and the test LSTM model root-mean-square-error is 0.69. There is a small difference between the train result and the test result, which is good. The Loss function for the Coca-Cola LSTM train model is shown below. From the graph, we can see the loss function is converging, which shows the LSTM model's convergence.

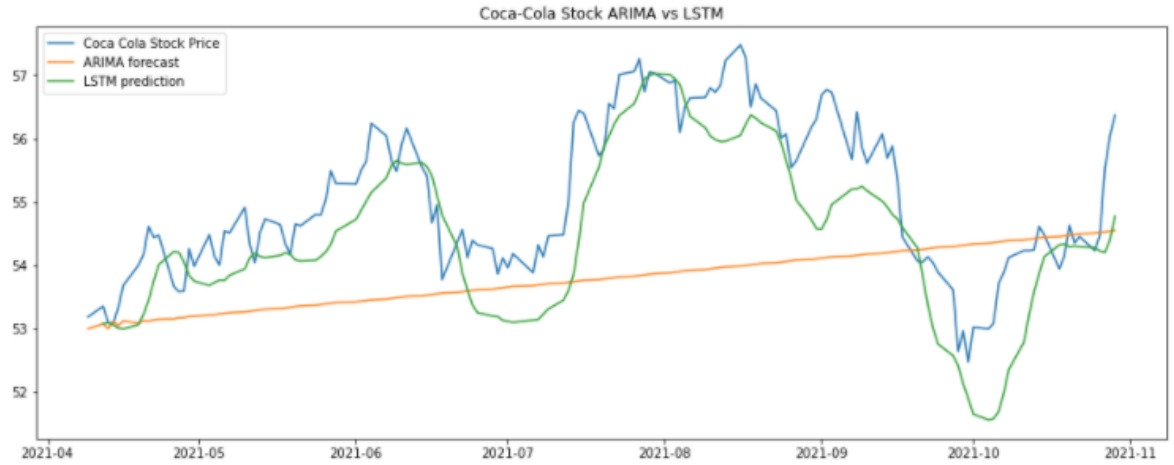


The out-of-sample prediction result of the test data of Coca-Cola stock is also quite good, it shows the basic trend of Coca-Cola stock.



Compare the out-of-sample performance between ARIMA Model and LSTM:





In the plot above, I put the LSTM prediction and ARIMA prediction for each stock in the same plot. From the plot, we can observe that there are more variation in the LSTM model. The prediction changes when the real stock price change. However, the ARIMA model prediction looks like a line which contains many small variation inside. For the ARIMA model, it focuses more on the long-term trend compare to LSTM model.

Focusing on ARIMA model, the pros and cons are quite obvious. The first advantage of ARIMA model is simple to implement and fast to run. However, there are many disadvantages for this model. There are a lot of limitations for ARIMA model. When predict time series, the time series must be stationary. In this way, when we use ARIMA model to predict stock price, we need to differ the data in order to make the data stationary. Essentially, it can only capture linear relationships, not non-linear relationships. From the plot, we can also find that the ARIMA model can only predict the long-trend of the stock price.

Compare to ARIMA model, LSTM seems a good choice when predict time series. The limitation of LSTM is quite low. There is no limitation on the data stationary and level shifts. Also, LSTM can model non-linear function with neutral networks. Moreover, since LSTM has memory capacity, it can remember more time step. However, the cons of LSTM stand out. The hardware requirements are very high for LSTM model. The LSTM model can deal with small-load data, when LSTM face very long-term time data, the process time will be very long.

In conclusion, when face the long-term stationary time-series, and want to predict the trend of the time series, we can use ARIMA model. If the time series we want to predict is not stationary and is non-linear, we can use the LSTM model to predict.