Amazon Stock Price Prediction

Vinesh Ramesh
PES1UG20CS504
Department of CSE,PESU University
vinnyveera03@gmail.com

Srishti Rayudu
PES1UG20CS329
Department of CSE,PESU University
rayudu.srishti7@gmail.com

***Abstract*—Amazon Inc. is one of the top e-commerce website in the world who is part of the Big Five American information technology companies alongside Alphabet, Apple, Meta, and Microsoft. The main objective of this project is to predict opening and closing price of the Amazon stock each day. By analysing the techniques implemented in previous published papers and the shortcomings of their results have been summarised in the following sections. Our main focus will be on the Long Short-Term memory (LSTM) model while also showing its benefits over other commonly used models like KNN,ARIMA and regressive models.**

Keywords—Amazon stocks, stock prices, predictive modelling, LSTM

I. Introduction

Amazon Inc. is of the leading e-commerce website whose expertise has diversified into new sectors such as cloud computing, digital streaming and artificial intelligence. Being one of the top reputed companies in the world, its stock also holds a lot of value. A stock, also known as equity, is a security that represents the ownership of a fraction of the issuing corporation. Amazon went public with its initial public offering(IPO) on May 15,1997 and has undergone four stock splits till date. To give context to how valuable the stock is, if you invested 10000 dollars on its opening day, it would be worth about 16 million dollars today. Therefore, one of the priorities for an investor is able to predict the future prices of the stock which would in return benefit them by making a ton of profit.

Stocks are majorly of two types, common and preferred. If you are a holder of a preferred stock, you get to exercise the voting rights in corporate decisions. Stocks hold great potential to investors for income growth over short or long term periods. Although, profits are not guaranteed all the time in the stock market, losses can be minimised by diversifying your portfolio into various companies. One such such kind of stock is blue chip stock which is the stock of huge company with huge reputation. Usually, their stock never loses value in the long term and is stable investment for investors that want to make profit over long term.Amazon is one such stock with a market cap of 1.22 trillion dollars and had a revenue of about 485 billion dollars in the year 2021-2022.

Our dataset contains information on the AMZN stock from the day of its opening, May 15,1997 to October 27,2021. Time series data will allow us to visualise all the data over fixed time intervals for us to make accurate measurements. We will use the parameters of open, close , low and high prices of each day for our study and implement different data visualisations to help choose the best model for our dataset.

II. Related work

We detail the methodologies and approaches used in other research papers in this section.

In this paper [1], the primary aim was to predict closing price of each stock on the next day and improving forecast accuracy. The approaches included linear and non-linear methods. Each time series was decomposed into component series using wavelet methods.

The paper focuses on what the mathematical concepts are that describe ARIMA models and LSTM networks, which are used to estimate future values of the chosen time series. With respect to wavelet methods, either individual or multivariate inputs could be used for the models.

The parameters for an ARIMA process is very important in order to get the most appropriate model to the time series. To estimate the number of Autoregressive (AR) terms, Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) can be used. Akaike Information Criterion (AIC) helps determine relative performance of a model with a different set of the above mentioned parameters.

Long Short-Term Memory (LSTM) Networks being a subclass of Recurrent Neural Networks (RNN), are suitable enough to learn long term dependencies in time series which a typical RNN fails to do as it is inclined for recent inputs or short term memory. With the introduction of gate functions, the phenomenon of long term memory is improved in LSTM. These gates are, Filter Gate, Memory Gate and Output Gate. These gates ensure transforming the state of current step to be used in the upcoming states. A 1 node and a 4 node LSTM hidden layer with both having an input layer of 3 nodes. In an Artificial Neural Network (ANN), each neuron receives more than one input followed by updating it current state and emit an output. A low weighted synapse means that output of the previous node is ignored, which is determined by the influence of it on the output neuron. Although, ANN has its limitations. One of them being its cyclic nature. Due to this particular feature, all inputs and independent of the outputs which is not ideal in the case of a time series model.

This drawback is addressed by RNN's feature of feedback loops with exist within its hidden layers. As a non orthogonal redundant transform, Maximal Overlap Discrete Wavelet Transform (MODWT), repeats information in neighbouring coefficients by overlapping time series values. In doing so, the cardinality of the original series and coefficient sets is maintained. Furthermore, another transform, Inverse MODWT, IMODWT reconstructs the original series. WAV-ARIMA, an extension to the ARIMA model, this method applies the MODWT to the series before applying the ARIMA model as above.

Closing prices of the stocks in FB, AAPL, AMZN, NFLX, GOOG from 1st January 2010 to 1st January 2017 are the main focus. Corporate actions that happened in this period influenced the number of observations for each stock. Google underwent a corporate reconstruction in March 2014

which is the centre of the analysis. These times series have been taken from Alpha Vantage. The models produce 100 one- day forecasts. PREVCLOSE is for measuring relative prediction improvements in other models. Predicted value, x_t is the previous day's closing price, denoted $\hat{x}_t = x_t - 1$. Similar to LSTM, WAV-LSTM additionally has a difference in the transformation of the series into the feature and target spaces, where the real values are replaced with vectors.
Multiple analyses were done based on the model implementations mentioned above. For each series, baseline forecasts were used to determine accuracies for all models for each time series. ARIMA was applied to the two LSTM topologies. 4-level for WAV-LSTM and 7-level Haar for MODWTs. Haar and db4 4-level and 7-level MODWTs for WAV-LSTM. Totally, there were 13 models for each 5 series.
Root Mean Squared Error (RMSE) was calculated as an accuracy measure as it has the same dimension or unit as the closing price of a stock. It was observed that, ARIMA and LSTM outperform PREVCLOSE with 20-30% accuracy. ARIMA outperforms LSTM. 1 node LSTM hidden layer outperformed the 4 node layer RMSE and MODWT RMSE. ARIMA outperformed its WAV-ARIMA versions. APPL 4 node WAV-:LSTM 4-level Haar had 2,86% increase in accuracy uplift and AMZN had 4.21% decrease.

By incorporating PCA, 4-level and 7-level decomposition could make a more informed decision. An ARIMA with addition explanatory variables (ARIMAX) or the General Autoregressive Conditional Heteroskedasticity (GARCH) model could better describe why ARIMA has constant variance.

In the paper [2], the aim was to predict spot price for a particular instance, say, region for a specific time in the future, and availability zone. Data used: the c3.2xlarge Linux instance type from the us-east-1b region between September 3, 2016 and September 10, 2016. LSTM was used as a main component of the RNN as they can identify and retain latent features over some time periods. It is important that the input data for the RNN is preprocessed into better format. Due to this, hyper parameter selection is important with 1 to 4 layers and 16 to 128 nodes for each layer. The one which gave the best result was three-layer solution with two LSTM layers and a dense layer for consolidation. These two layers are each 32 units wide. The LSTM units can accept data sequentially. As LSTM has three gates which enable to forward or forget data, it is preferred over traditional RNN. ADAM optimisation model was used to train the neural model along with Nesterov momentum and Mean Square Error as the loss function.
For the preprocessing stage, there are two methods namely, method to regularise pricing data relative to the on-demand price, and the method that applies exponential smoothing to regularise data for seasonal patterns. There is a better expansion from one instance type in one region to all instance types in all regions by transforming raw data into a percentage of the on-demand price. It is much less computationally expensive to retrain a network from values between 0.1 and 0.3 to values between 0.5 and 0.7 than it is to retrain a network from values around 10 to values around 0.5. The next step in preprocessing is to use Holt-Winters exponential smoothing to get a well-represented approach for removing the easily quantifiable trends so that the more focus is on predicting the irregular trends of the time series. By doing so, the does not have to learn all the broad trends. Amazon spot pricing is used to evaluate the network and assess its performance against a baseline model, ARIMA. This baseline model has proven to perform well for spot price prediction. MSE is the primary metric to evaluate. Furthermore, Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are also recorded.
10,000 points were for training and 2.000 were for validation. Given 50 input observations, forecast was to be done for the subsequent 50 periods as 50 observations meant a window of 1 hour. The model was trained 10,000 points 250 epochs which was applied on the 50 input validation points. Spikes are common with ADAM and are potentially exaggerated via NADAM's Nesterov (N) momentum component. Training MSE is 7×10^{-6} which was observed without preprocessing. The MSE over the 2,000 point validation set was 1.7×10^{-5} for the neural network and 4.2×10^{-5} for ARIMA. The Mean Absolute Percentage Error (MAPE) for the ARIMA and LSTM models were 3.76% and 3.35%, respectively. Here. The LSTM is more accurate than the ARIMA model. RMSE for ARIMA is 0.557 and 0.423 for LSTM. Whereas, MSE for ARIMA is 0.362 and 0.320 for LSTM. A perfect prediction would have its MSE equal to 0. Now with preprocessing, using Holt-Winters exponential smoothing, a training MSE of 6.98×10^{-6} was observed which indicates a seasonal trend in the data can be detected by the neural network.

Another paper [3] covers stock market prediction based on Deep Long Short Term Memory Neural Network. Stock Vector is basically a word vector whose dimension is reduced and then expressed in a low dimension space. This helps on forecasting the stock market. The Deep Long Short Term Memory Neural Network with Embedded Layer based on Stock Vector (ELSTM) which is an improved RNN is used here. The gap is too big as the opening price of shares is different. An embedded layer is added to the LSTM for more precise stock prediction. This paper used Theano framework on Python as its software environment and Centos 64 bit OS, 62G RAM, Intel(R) Xeon(R) CPU as its hardware environment. In order to get information about the stock, crawler technology has been used. All data about the stock is filtered based on five indicators, lowest price, highest price, closing price, opening price and volume of the stock. This helps in predicting the price and trend of the Shanghai A-share Composite Index. A few additional indicators are added as well. For this particular study, the chosen stocks are from 2006/1/1 to 2016/10/19 of which 70% is divided into training, 10% for validation and the rest is for testing. Furthermore, the data is normalised on [0,1]. This paper evaluates the performance of the model for short-term stick price forecast from two aspects, Mean Squared Error (MSE) and Accuracy.
First LSTM was used, then a better ELSTM model was used for selected stocks. Average accuracy was 53.2% and A-share Composite Index is 57%. This is higher than stochastic forecast. In conclusion, the mentioned methods perform better for Shanghai A-share Composite Index.

Another paper [4] that we came across had some interesting insights. Data was obtained from years between 2014 and 2019 where stocks of Amazon, Apple, Google and Microsoft were considered. The primary objective was to get an estimation of the near-term stock fluctuation. There are three classes that have been considered in this paper, -1, 0 and 1. Volatility and momentum being important aspects of the stock market have been incorporated in the process of estimation. However, in terms of predicting the direct in which the stock market is moving if proves to be robust. The final accuracies for all the considered data ranges between

90 and 95% for a 7-day window. In order to perform statistical analysis, KS test and KL divergence test have been employed. The goodness of the fit for all four stocks have been separately recorded. Several methods show the robustness of the LSTM network, that is, p-values that have been obtained from the KS test and the entropy derived from the KL divergence test.

Finally, [5] focused mainly on how artificial intelligence and machine learning algorithms can could help for future stock market predictions. In this paper, Google, Facebook, Apple and Amazon were the stocks taken into consideration. Data the study was mostly taken from yahoo finance. As far as Amazon stocks were concerned, data was picked between January 2019 and July 2019 which was first normalized. Several methods such as, Linear Regression, Prediction using 3-month average measurements, prediction using Exponential Smoothing were used. Upon comparing all results, it was concluded that exponential smoothing prediction resulted in the least error and greater accuracy making it the best stock market predictor.

III. PROBLEM STATEMENT

The dataset chosen for this project is sourced from kaggle and is linked below.
https://www.kaggle.com/datasets/kannan1314/amazon-stock-price-all-time
The data recorded is for about the span of 24 years, from 15 May 1997 to 27 October 2021.Our job is to analyse the provided data and predict the opening and closing price of the Amazon stock for future dates.

A. Dataset

The columns present in the above dataset are date, open, close ,high, low, adjacent close and the volumes of trades done in a particular day.
'Date' attribute indicates the dates for which relevant information is present. 'Open' indicates the opening price of the stock during the start of the trading session. 'Close' indicates the closing price of the stock during the end of the trading session. The 'High' attribute refers to highest price the stock is traded at during the session. The 'Low' attribute refers to lowest price the stock is traded at during the session. The 'Adj Close' attribute refers to amended price of the stock's closing price after accounting. 'Volume' attribute refers to the quantity of shares of stocks traded during the session.

B. Exploratory Data Analysis and Visualisations

In our dataset, there are 7 attributes with 6155 rows each.We notice the date attribute is object while the rest of the attributes are float64 while volume is an int64 datatype.We figure out the mean, standard deviation, minimum, maximum and percentiles of each attribute in our exploratory data analysis. There are neither duplicated nor empty values in our dataset.
We also infer that dataset contains a lot of outliers in our dataset. The normal outliers lie in the range of 950s for each attribute whereas the extreme outliers lie in the range of 420s. They are not removed as they hold significant value for our further analysis. The correlation plot(fig 1) reveals that there is very large positive correlation between all attributes other than the volume column. Volume refers to the amount of shares traded in a day whereas rest of the attributes describe the price of the stock.Therefore, there is negative correlation between volume and other columns.

	Open	High	Low	Close	Adj Close	Volume
Open	1.000000	0.999934	0.999912	0.999842	0.999842	-0.240087
High	0.999934	1.000000	0.999893	0.999924	0.999924	-0.238907
Low	0.999912	0.999893	1.000000	0.999928	0.999928	-0.241351
Close	0.999842	0.999924	0.999928	1.000000	1.000000	-0.240122
Adj Close	0.999842	0.999924	0.999928	1.000000	1.000000	-0.240122
Volume	-0.240087	-0.238907	-0.241351	-0.240122	-0.240122	1.000000

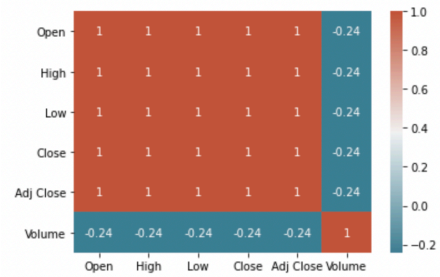


Figure 1.

In figure 2, we notice that there is significant amount of outliers present in the scatter. This is common in real-time stock data and holds great value for further processes.

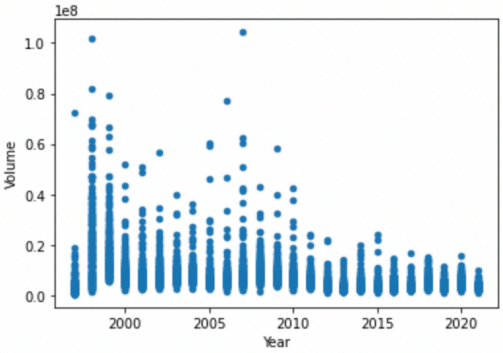


Figure 2.

In figure 3, we see the opening prices and closing of stocks prices for the first sixty values. As you can see, it is very common to see that the opening prices are usually greater than closing prices for each day.After hours trading (AHT) has had a major effect on the price of the stock between the closing and opening bells. Therefore, we don't see the opening price of stock on a day equal to its closing price on the previous day.

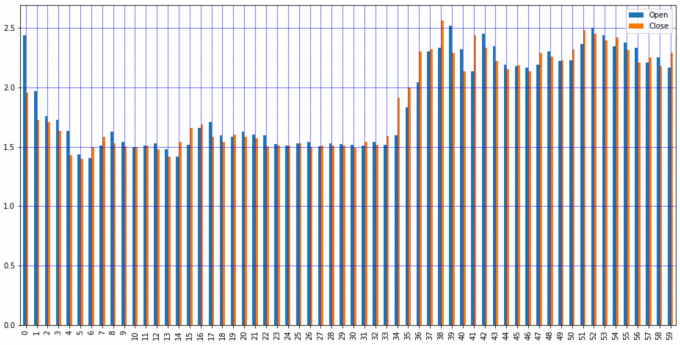


Figure 3.

In figure 4, we can see the opening stock prices for each day from 1997 to 2021. In initial phases, the price of the stock did not fluctuate much and was steady. Recently from 2015, the price has taken a steep rise and is still rapidly growing reaching figure of over 3500 dollars.

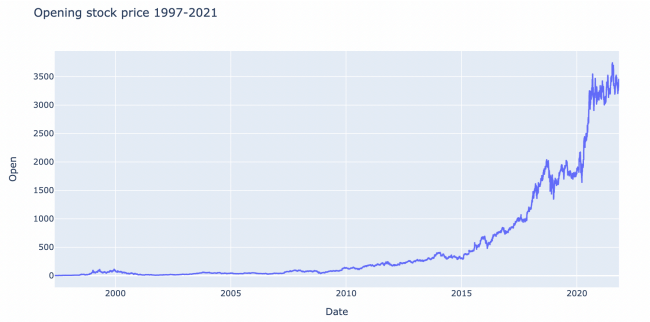


Figure 4.

In figure 5, we observe the amount of stocks that we traded each day. We see distinct spikes during the years 1998 and 2008. The 2008 spike was caused due to a great recession and financial crisis. Due to the losses occurred in this time period, people began to lose faith in the stock and eventually the amount of stocks traded decreased in number.

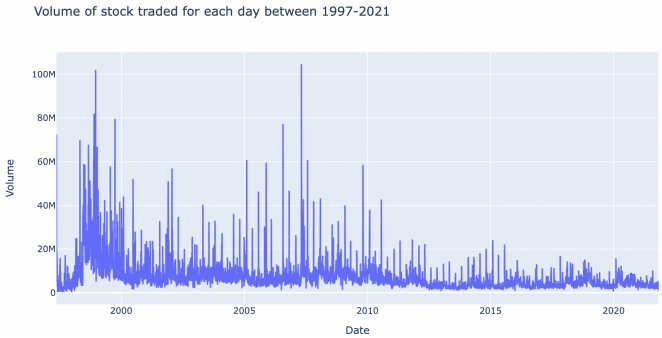


Figure 5.

In figure 6, we obtain the normalised value of the stock prices adjusted throughout the years. We can really see that the value of amazon stock has grown substantially looking at the value at the top of the table compared to the bottom of the table. Its value has at least 3300x over 24 years of its existence.

	Open	High	Low	Close	Adj Close	Volume
0	0.000276	0.000279	0.000166	0.000151	0.000151	0.690172
1	0.000150	0.000141	0.000107	0.000089	0.000089	0.136869
2	0.000095	0.000086	0.000085	0.000084	0.000084	0.054117
3	0.000086	0.000080	0.000087	0.000064	0.000064	0.047957
4	0.000061	0.000052	0.000017	0.000008	0.000008	0.176865
...	...	...	...	...	...	...
6150	0.911893	0.911762	0.920500	0.920537	0.920537	0.013426
6151	0.913696	0.908994	0.901098	0.893872	0.893872	0.025487
6152	0.890718	0.887242	0.892006	0.889802	0.889802	0.016745
6153	0.894594	0.905357	0.904529	0.904735	0.904735	0.021249
6154	0.904879	0.904264	0.911964	0.910129	0.910129	0.005711

6155 rows × 6 columns

Figure 6.

In figure 7, we see the pairwise scatter plot for all attributes in our dataset. It further solidifies the strong correlation between all other attributes other than volume as discussed before.

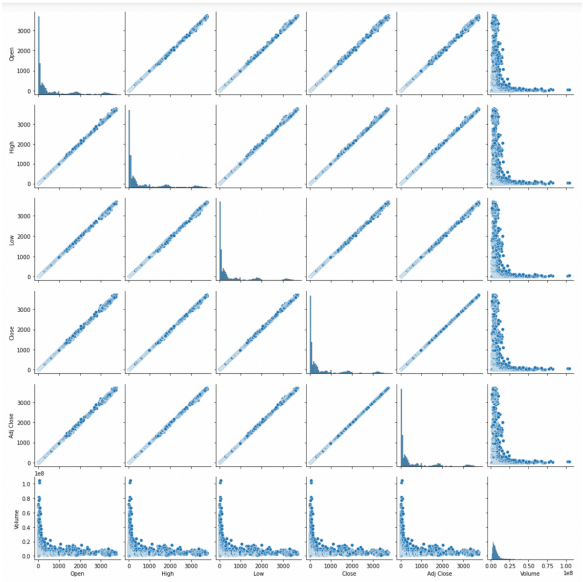


Figure 7.

Figure 8 shows us the time series decomposed.

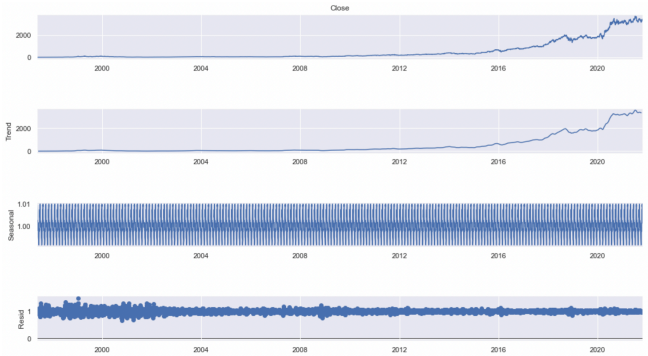
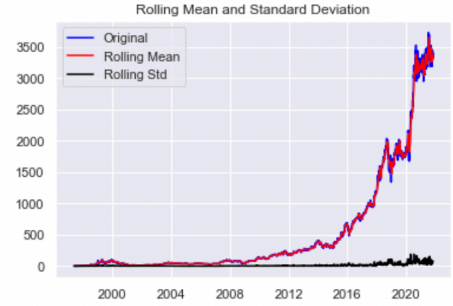


Figure 8.

As you can see in Figure 9 , we get p-value of 1.00 which is greater than 0.05 which shows us that the time series is non-stationary.




```
Results of dickey fuller test
Test Statistics                  3.912500
p-value                          1.000000
No. of lags used                34.000000
Number of observations used    6120.000000
critical value (1%)              -3.431419
critical value (5%)              -2.862012
critical value (10%)             -2.567021
dtype: float64
```

Figure 9.

IV. PROPOSED SOLUTION

After careful study of the above mentioned papers and exploring different models, we chose the LSTM as the starting point of our analysis. We will also compare the working of different models like ARIMA and regression and further our analysis depending on the results. For now, our focus is building the LSTM model and refining it to best fit our dataset.

A.Multiple Linear Regression(MLR)

Multiple linear regression is establish a linear relationship between the target variable and explanatory variable.Target variable is typically the attribute you want to predict whereas the explanatory variables are used to explain the previous tern and try to extrapolate the existing trend into the future. We choose the 'Close' column as the target variable and 'Open', 'High' and 'Low' variables as the explanatory variables. There exists a strong correlation between the variables as seen using the heat-map. The dataset was split as 80:20 and was used to predict closing prices of stocks of each day.

B.ARIMA

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. Its primary goal is to predict the future stock prices by analysing the difference between the values in series. To deploy ARIMA model, we need to make use our data is stationary. It is an effective and diverse model which gives us relatively accurate answers. We choose the 'Close' column as the target variable and 'Open', 'High' and 'Low' variables as the explanatory variables.

C.LSTM

LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. It only remembers the relevant data and forgets the non essential features.Information can be added to or removed from the cell state in LSTM and is regulated by gates.
We choose to use univariate and multivariate LSTM with sliding window of 30. Univariate LSTM uses a sliding window of 30 data points and its input consists of only the last 30 'Close' column values to predict the 31st value. The latter i.e., Multivariate LSTM also uses a sliding window of 30 data points but here, the input consists of the last 30 'Open', 'High' and 'Low' columns' values to predict the 31st value of the 'Close' column.The dataset is split into training and test data(80-20) and was scaled using the MinMaxScaler that transforms features by scaling each feature to a given range.

V. EXPERIMENTAL RESULTS

A.LSTM

The univariate LSTM was trained on 20 epochs. With a loss of 1.2516e-04 and an RMSE of 405.4695208901257.Even though the curves look similar, it can not anticipate the sudden rise in the value of the stock.

Figure 10.

The multivariate LSTM trained on 20 epochs with loss of 1.0145e-04 and RMSE of 1272.4697783047036. The model is pretty much useless as seen below.

Figure 11.

B.Multiple linear regression(MLR)

With a prediction score of 0.999923769834731 and RMSE of 7.344728268189395, we find the MLR model to be way more suited to our data than LSTM. We can view relevant information below.

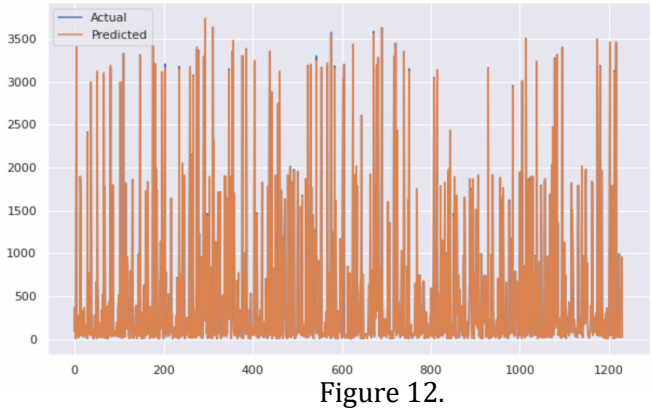


Figure 12.

C. ARIMAX

Using the auto_arima method, with the exogenous factor as Volume, we were able to obtain optimal values of 0,1,0 for p, d and q respectively. We achieved an excellent RMSE of 0.9739560696471813 and results are shown below.

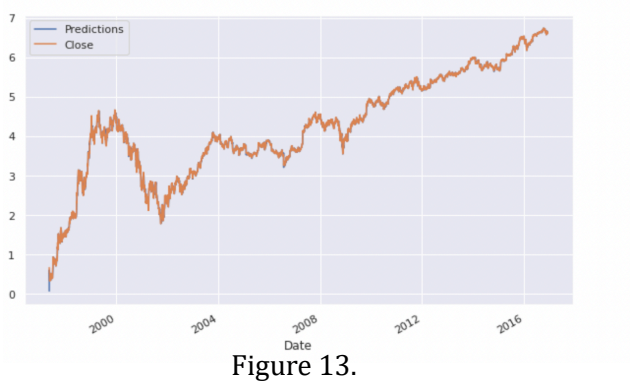


Figure 13.

D. ARIMA

Using the auto_arima method, with the exogenous factor as Volume, we were able to obtain optimal values of 0,1,0 for p, d and q respectively. We achieved a slightly worse RMSE of 0.9820668852593264 and results are shown below.

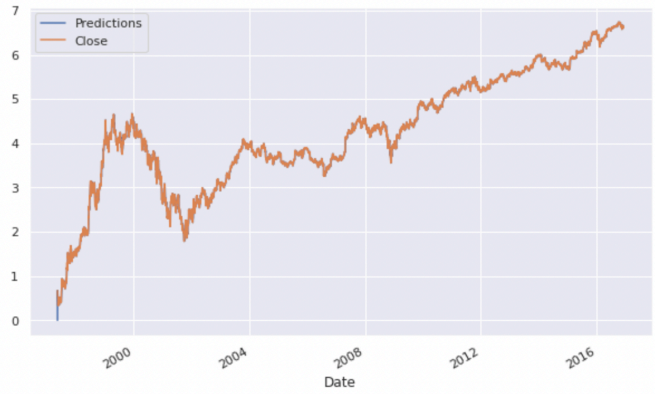


Figure 14.

VI. CONCLUSIONS

After trying out various models like LSTM, ARIMA and MLR , we choose the best model according to the RMSE value. Ranking our models from best to worst we have , ARIMAX, ARIMA , MLR, LSTM univariate and LSTM multivariate.
We can try to achieve more accuracy by trying different exogenous feature for ARIMAX model and other values of p, d and q. We can also implement deep learning algorithms with tensorflow to try models like RNN.
As of now , we were able to predict the closing prices of the stocks most accurately using the ARIMAX model.

VII. REFERENCES

[1] Tom Skehin, Martin Crane, Marija Bezbradica “Day Ahead Forecasting of FAANG Stocks Using ARIMA, LSTM Networks and Wavelets.”

[2] Matt Baughman, Christian Haas, Rich Wolski, Ian Foster, Kyle Chard “Predicting Amazon Spot Prices with LSTM Networks”

[3] Weiwei Lin, Pan Wang, Yanqiang Zhou , Xiongwen Pang “Stock Market Prediction based on Deep Long Short Term Memory Neural Network” 3rd International Conference on Complexity, Future Information Systems and Risk.

[4] Shubham Ekapure, Nuruddin Jiruwala, Sohan Patnaik, Indranil SenGupta “A data-science-driven short-term analysis of Amazon, Apple, Google, and Microsoft stocks”

[5] M Umer Ghania, M Awaisa, Muhammad Muzammula “Stock Market Prediction Using Machine Learning(ML) Algorithms”