International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

Stock Market Analysis a Naive Approach to Solve it Using Data Science

Shagil Chaudhary [1], Shouvik Dasgupta [2]

U.G. Student, Department of Computer Engineering, Jamia Hamdard University, New Delhi, Delhi, India[1]

U.G. Student, Department of Computer Engineering, Jamia Hamdard University, New Delhi, Delhi, India[2]

ABSTRACT: In such a volatile environment, scaling a business is hard. It takes considerable effort. Dealing with sales and marketing, understanding taxes and corporate compliance, interacting with customers on a daily basis delivering on any commitment to them and so much more. At the end of the day, companies that manage to control all of this grow consistently while others have to struggle. In this project, we have tried to forecast which company will grow in the near future and which will struggle to survive. We have used the New York Stock Exchange dataset (https://www.kaggle.com/dgawlik/nyse) for the analysis. But how are we going to analyze the fate of the companies given only their stock prices?

KEYWORDS: Time Series, Stock Prediction, Forecasting.

I. INTRODUCTION

Given the stock exchange values of various companies, we will try to search for the company having the most overvalued and undervalued stock.
An overvalued stock has a current price that is not justified by its earnings outlook or price/earnings (P/E) ratio, so it is expected to drop in price. Overvaluation mostly results from a deterioration in a company's financial strength.
We will take the most overvalued stock, and predict whether it will go bankrupt or at least lose its value in the near future. The same goes for the undervalued stock.
Undervalued or overvalued stock can be identified using their P/E ratio. A low P/E can indicate either that a company may currently be undervalued or that the company is doing exceptionally well relative to its past trends. Hence it has a chance of doing well in the future as well. A high P/E value means the stock is overvalued and will go down in the near future.
But what Is Price-to-Earnings Ratio – P/E Ratio?
The price-to-earnings ratio (P/E ratio) is the ratio for valuing a company that measures its current share price relative to its per-share earnings (EPS). The price-to-earnings ratio is also sometimes known as the price multiple or the earnings multiple.
Formula:

$$\text{P/E Ratio} = \frac{\text{Market value per share}}{\text{Earnings per share}}$$

In simple words, the P/E ratio is calculated by dividing a company's current stock price by its earnings per share (EPS). EPS is calculated by subtracting a company's preferred dividends paid from its net income and then dividing the result by the number of shares outstanding.
For both, the overvalued and the undervalued stock, we will forecast their stock's closing price and ensure whether the company will grow or struggle in the upcoming years.
Assumptions –
As this forecasting is relative,
We have considered the highest price as their current stock price.
We have considered the difference in the companies closing and opening price over the past few years as their income.
Volume is taken as the number of shares outstanding.

II. PRELIMINARY ANALYSIS OF THE DATASET USING TABLEAU

The dataset contains about 501 stocks in total. It was illogical to perform forecasting and any kind of analysis on all of the stocks so a sample of the complete dataset is taken. The selection is done on the basis of the closing amount of the

stocks in regards to the EPS and PE ratio. In order to make the process simple and more efficient tableau is used for preliminary data analysis and sorting.

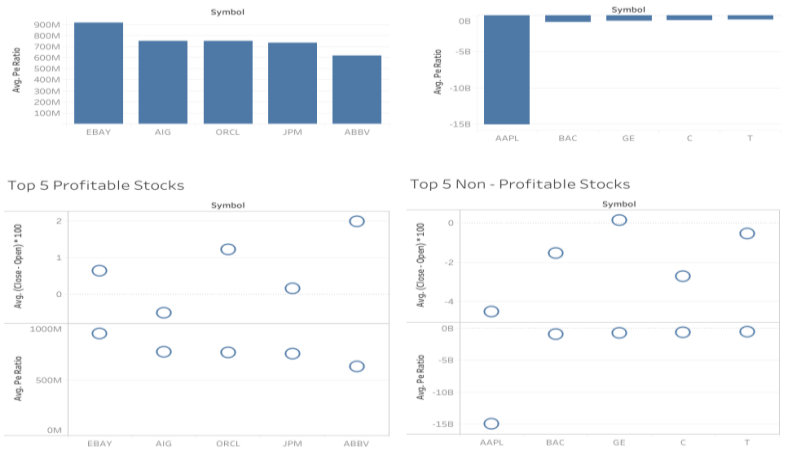

Figure - 1 (profitable and non-profitable stocks)

Ebay has the most non-profitable stock on an average and Apple has the most profitable stock on an average. Hence Ebay's and Apple's stocks are selected for forecasting.

III. ALGORITHMS USED FOR FORECASTING/PREDICTING STOCK PRICES

A. Facebook Prophet Time Series

Facebook developed an open-source forecasting tool named "Facebook Prophet". It provides various functionality by hyper tuning its parameters. The algorithm is based on a decomposable time series model.

The decomposable model consists of three main components.

1. Trend
2. Seasonality
3. Holidays

The combining equation holding together all the components can be given as :

$$y(t)= g(t) + s(t) + h(t) + \varepsilon t \qquad (1.1)$$

1. g(t): for modeling non-periodic changes in time series
2. s(t): periodic changes
3. h(t): effects of holidays with irregular schedules
4. εt: for any unusual changes not accommodated by the model

For Ebay's stock the rmse value comes out to be approximately 3 and the mape value is also good so we can safely presume that the forecast done for the ebay stock can be considered. According to the forecast it is evident that although the stocks are in profit but are likely to go up a few points only.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

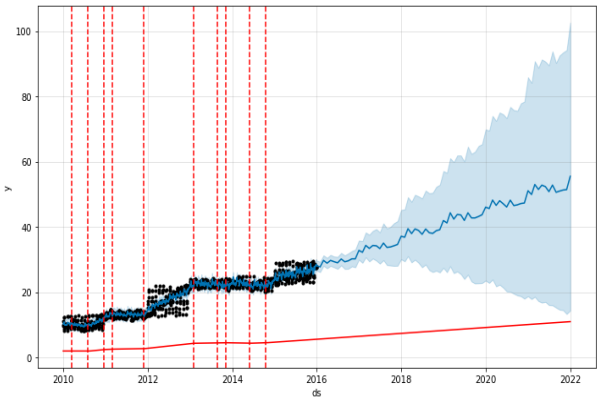

Figure - 2 Ebay stock prediction

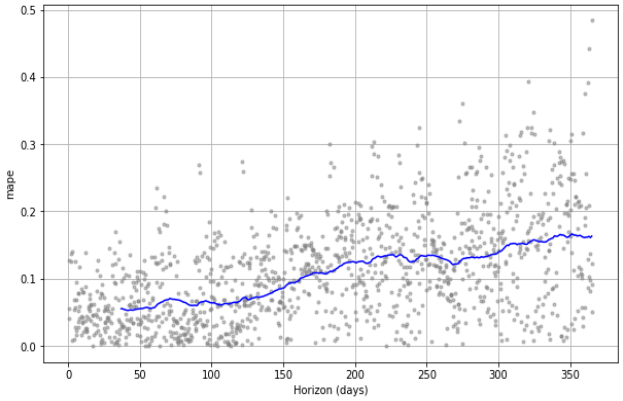

Figure - 3 Algorithm MAPE value spread

For Apple's stock the rmse value comes out to be approximately 17 and the mape value is also good so we can safely presume that the forecast done for the apple stock can be considered. According to the forecast it is evident that the stock is definitely in profit and is likely to go up a few points.

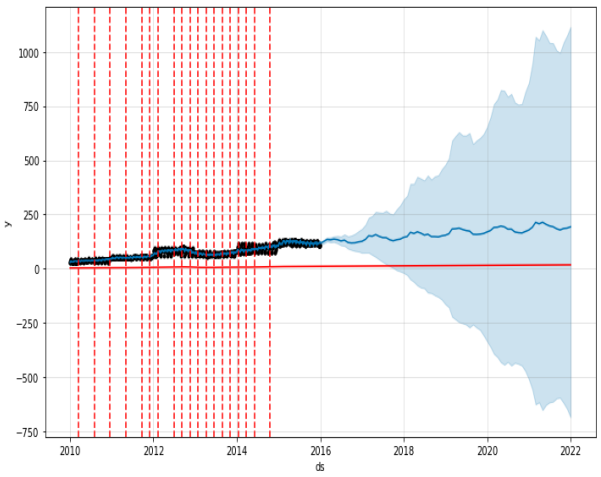

Figure - 4 Apple stock prediction

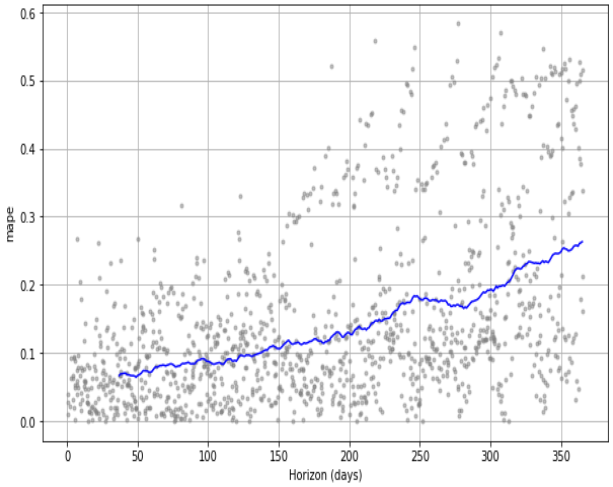

Figure - 5 Algorithm MAPE value spread

B. Simple Moving Average

The algorithm of a simple moving average is a simple technique that smoothes out the data points by the creation of a constantly updated average price. This average takes place over a period of a specified timeframe. Moving average strategies are also popular and can be tailored to any time frame, suiting both long-term investors and short-term traders. In the moving average algorithm we will take into consideration the closing values of the stock so as to predict the future value.

While calculating successive values, a new value comes out and the oldest value is dropped, so a full summation each time is unnecessary.

$$\bar{p}_{\mathrm{SM}} = \bar{p}_{\mathrm{SM,prev}} + \frac{1}{n}(p_M - p_{M-n}). \tag{1.2}$$

With the moving average the rmse value comes out to be approximately 12 which points out that the prediction of the value was accurate. It is evident that the value of the stock is expected to go up in the near future.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

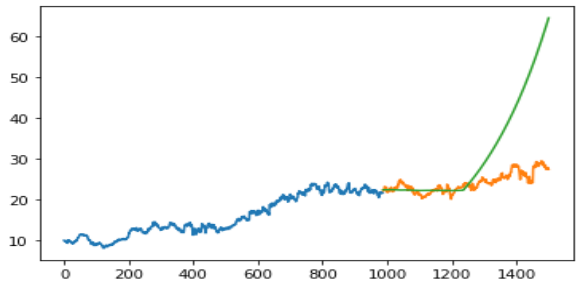

Figure - 6 Ebay stock value prediction

There are some anomalies with the stock dataset that could not be removed which may or may not have an effect on the prediction. With the moving average the rmse value comes out to be approximately 34 which points out that the prediction of the value was somewhat accurate. It is evident that the value of the stock is expected to go up in the near future.

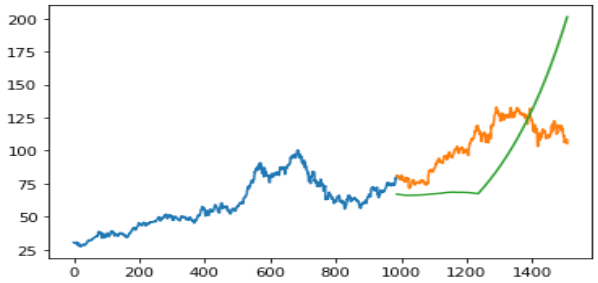

Figure - 7 Apple stock value prediction

C. Auto ARIMA (Auto-Regressive Integrated Moving Average)

ARIMA is a popular statistical method for time series forecasting. For ARIMA o work the data series needs to be stationary i.e. the variance should not vary with time.

$$ARMA(p',q) = X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad (1.3)$$

For better results, the algorithm can be hyper tuned. The tuning done is as follows:

1. Taking into consideration, the season data.
2. The starting value of p and q stepwise is provided with the maximum value p and q can go.
3. Trace is stated true which will give a list of ARIMA models considered.

With the Auto ARIMA algorithm, the rmse value turns out to be approximately 1 with means that the prediction of the value was accurate. According to the prediction Ebay's stock prices as likely to go up.

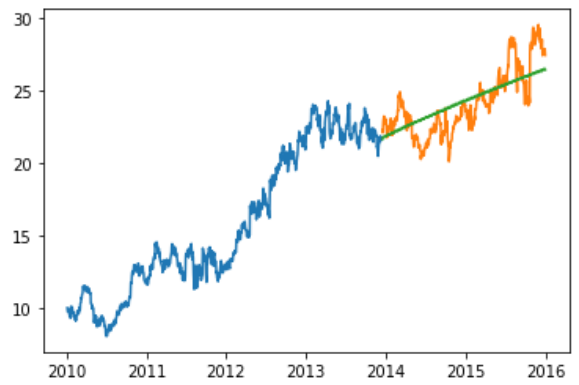

Figure - 8 Stock value prediction

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

With the Auto ARIMA algorithm the rmse value turns out to be approximately 24 with means that the prediction of the value was accurate. According to the prediction Apple's stock prices as likely to go up and then fall a few points.

D. SARIMA (Seasonal Auto-regressive Integrated Moving Average)

Seasonal ARIMA or SARIMA is an extension of the underlying ARIMA algorithm. SARIMA has the ability to support univariate time series data which contains a seasonal component.

SARIMA(p,d,q)(P,D,Q) is given by

$$\Phi_P(B^s)\varphi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t \qquad (1.4)$$

Where the variables can be described as:

p for Auto-regressive part as to how many past values (p) our model is dependent on.

q for Moving Average part as how many past random error terms (q) our model is dependent on.

d for the Integrated part as the number of times differencing is required to make it stationary.

The usual characteristics that are found in a Time series data are :

1. Trend - If the values show any overall upward or downward trend or slope.
2. Seasonality - If the values repeat after a specific interval
3. Noise - no relations between the values and the time.

Given the usage of the most basic model, the actual and the predicted values don't seem to be far apart.

The RMSE value using the best parameters was - 2.192. Hence the Ebay's stock value will go up and then it will fall a few points.

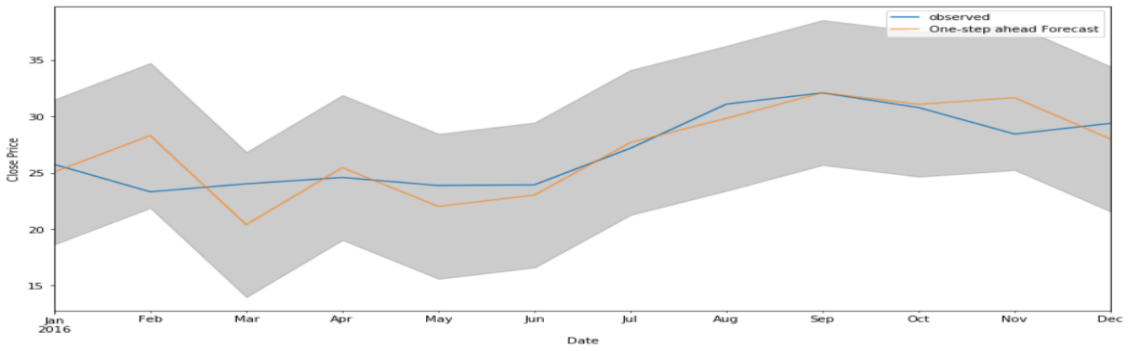

Figure - 9 observed vs forecast value (Ebay stock)

The RMSE value using the best parameters was - 22.1464. Hence the Apple's stock value will go up and then it will fall a few points.

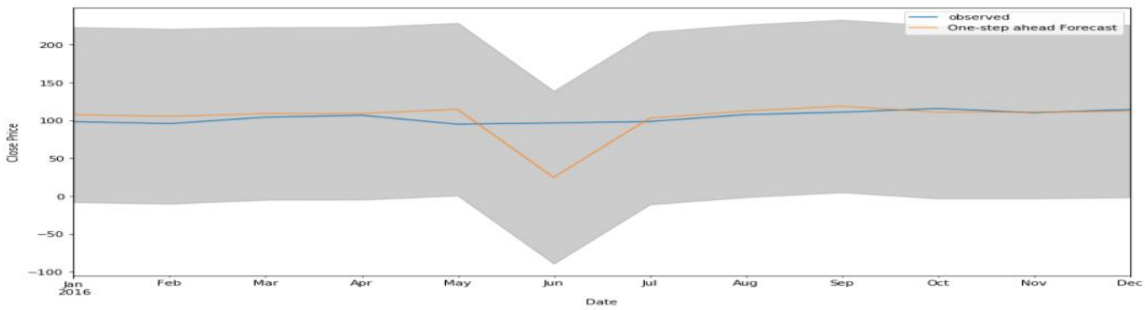

Figure - 10 observed vs predicted value (Apple stock)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

E. Linear Regression

Linear Regression is a basic algorithm that represents a relationship between a scalar and an explanatory variable.

The single explanatory variable category is called simple linear regression.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.5}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Taking into account the rmse value of approximately 4 we can safely presume that the prediction done for the ebay stock can be considered accurate. According to the prediction the stock value will climb up.

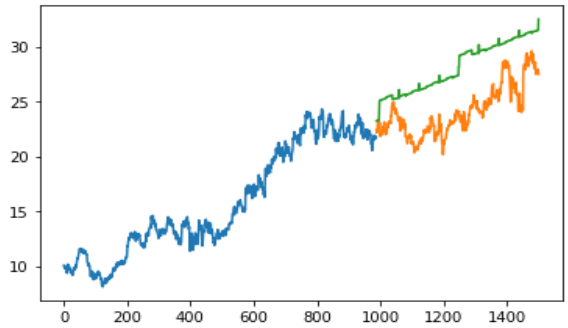

Figure -11 predicted vs actual value (Ebay stock)

Taking into account the rmse value of approximately 18 we can safely presume that the prediction done for the apple stock can be considered accurate. According to the prediction the stock value will climb up conclusively.

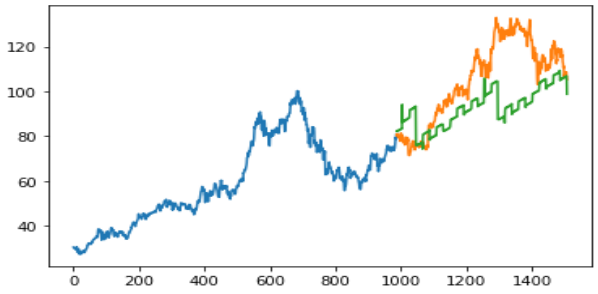

Figure - 12 predicted vs actual value (Apple stock)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

F. XG Boost

In order to increase the speed and performance of the regression algorithms gradient boosted decision trees are used resulting in an algorithm called XGBoost. This algorithm is popularly used for structured or tabular data. XGBoost is a function of functions and can be defined as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Real value (label) known from the training data-set

Can be seen as f(x + Δx) where x = $\hat{y_i}^{(t-1)}$

The model seems to overfit in most of the cases (which is sometimes common in XGBoost).

The test scores are decent but not up to the mark. Future work can be related to reducing the test RMSE value further by detailed fine-tuning the parameters and avoid overfitting the model. For the ebay stock prediction the MSE value was 1.20 and the RMSE value was -17.627. We can somewhat say on the basis of prediction that the value will go up.

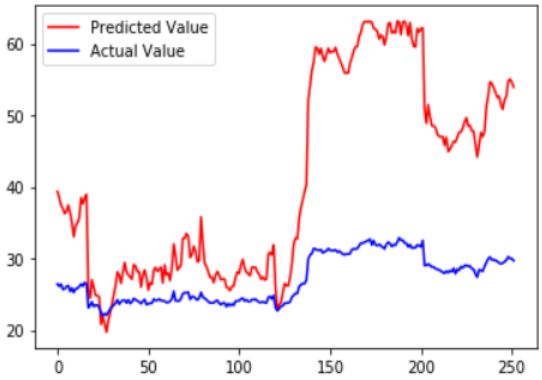

Figure - 13 predicted vs actual value (Ebay stock)

For the apple stock prediction the MSE value was -248.462 and the RMSE value was -337.6251. We can't make any concrete predictions in this case.

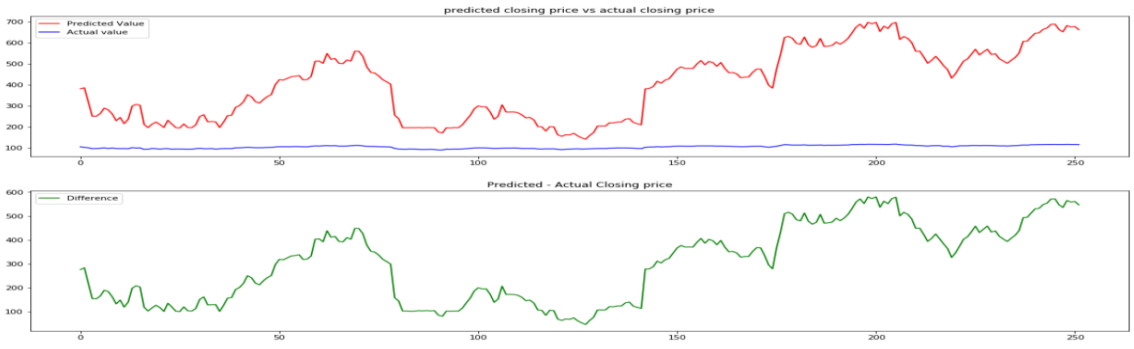

Figure - 14 predicted vs actual value (Apple stock)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

G. LSTM (Long Short-Term Memory)

LSTM is an algorithm having a recurrent neural network type that is capable of learning order dependence in sequence prediction problems.

The LSTM cell consists of the following components:

1. Forget gate
2. Candidate layer
3. Input gate
4. Output gate
5. Hidden state
6. Memory state

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

The LSTM works best having the following architecture:

1. GRU(256 neurons) with a dropout value of 0.4
2. LSTM(256 neurons) with a dropout value of 0.4
3. Dense layer

With an Adam optimizer (initial learning rate of 0.001), 150 epochs and a batch size of 128, the best score was -

Train RMSE of 1.17 and Test RMSE of 16.74.

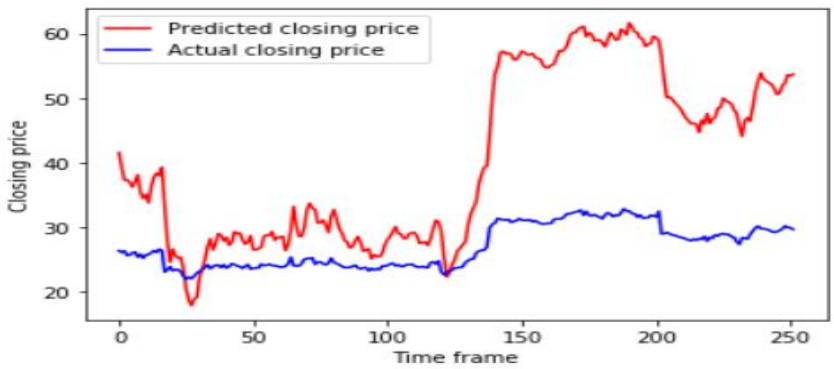

Figure - 15 predicted vs actual value (Ebay stock)

According to the prediction the ebay stock value may increase but the scores are not satisfactory so a concrete prediction can not be stated.

For the apple stock data the algorithm with an Adam optimizer (initial learning rate of 0.001), 150 epochs and a batch size of 128, the best score we got was -

Train RMSE of 8.37 and Test RMSE of 332.94.

The test scores are very poor and the model seems to be overfitting. Future work can be related to reducing the test RMSE value further by trying out other architectures and overcome the issue of overfitting. Hence a concrete prediction can not be provided.

Figure - 16 predicted vs actual value (Apple stock)

IV. EXPERIMENTAL RESULTS

Algorithms	Ebay's Stock RMSE	Apple's Stock RMSE
Facebook prophet time series	3.100	17.237
Simple moving average	12.719	34.697
Auto ARIMA	1.476	24.653
SARIMA	2.192	22.146
Linear regression	4.244	18.782
XGBoost	17.627	337.625
Long short term memory	16.74	332.94

Table - 1 Metric score of all the algorithms used

V. CONCLUSION

Plots forecasting the rough estimate of the closing stock prices for both Ebay and APPLE over the next few years –

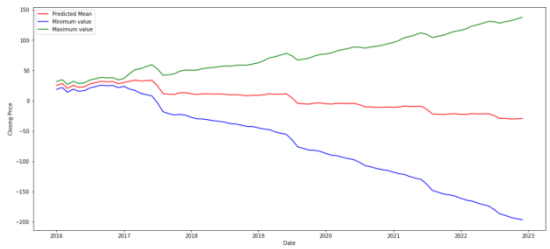

Figure -17 (Ebay's stock prediction)

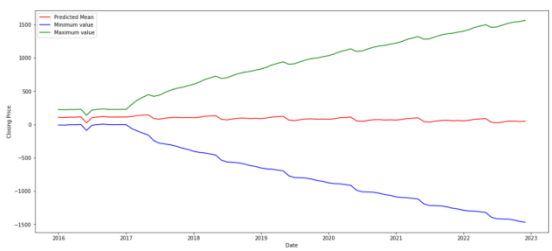

Figure - 18(Apple's stock prediction)

Our forecasts validate the initial findings which claimed Ebay to be the least profitable stock and Apple to be the most profitable one. Hence, we can say that Ebay will most probably lose its value over the next few years whereas Apple will shine over Wall street.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

This approach will surely assist the share market holders and traders to choose the most profitable stock in a longer span of time. The availability of company details would have further confirmed the authenticity of this approach. This approach, for now, is very basic with some assumptions as the comparison is relative.

Further work consists of –

1. Applying this approach to more authentic and reliable data
2. Fine-tuning the model further to get the best possible result

REFERENCES

[1]. An End-to-End Project on Time Series Analysis and Forecasting with Python. By Susan Li.
[2]. Price-to-Earnings Ratio – P/E Ratio. By Adam Hayes.
[3]. Forecasting stocks with Facebook Prophet. By Yibin Ng
[4]. Top 6 Regression Algorithms Used In Data Mining And Their Applications In Industry. By Richa Bhatia
[5]. Data Visualisation with Tableau. By Parul Pandey.