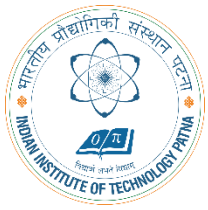


Indian Institute of Technology Patna

Department of Computer Science Engineering

Course

Applied Time Series Analysis
(CS575)

Mini Project

On

Analysis of financial time series

Submitted by

Vipin Gupta
Roll no- 2011MT22

Instructed by

Dr Jimson Mathew
Associate Professor
&
Dr Mayank Agarwal
Assistant Professor

Project Objectives

Analysis of financial time series

Consider a time series stock values of various companies for a given duration (download using the script below). Your task is to find an appropriate model for the stock prediction. Estimate the parameters of the chosen time series model. Compare at least with three different models. Elaborate on difficulties and alternative approaches.

Methodology and Mathematical Background

1. ARIMA (p, d, q) model

Its also called auto regressive integrated moving average model. It's made by 3 different word having 3 different argument parameters. It's "stochastic" modeling approach that can be used to calculate the probability of a future value lying between two specified limits.

1) Auto regressive- measure of no of lag observations p

AR(1) $y_t = a1 * y_{t-1}$
AR(2) $y_t = a1 * y_{t-1} + a2 * y_{t-2}$
AR(3) $y_t = a1 * y_{t-1} + a2 * y_{t-2} + a3 * y_{t-2}$

2) Integrated- measure of degree of differencing d

MA(1) $\varepsilon_t = b1 * \varepsilon_{t-1}$
MA(2) $\varepsilon_t = b1 * \varepsilon_{t-1} + b2 * \varepsilon_{t-2}$
MA(3) $\varepsilon_t = b1 * \varepsilon_{t-1} + b2 * \varepsilon_{t-2} + b3 * \varepsilon_{t-3}$

3) Moving average- size/width of the moving average window q
To build a time series model issuing ARIMA, we need to study the time series and identify p, d, and q
   - Ensuring Stationarity- Determine the appropriate values of d
   - Identification- Determine the appropriate values of p & q using the ACF, PACF, and unit root tests. p is the AR order, d is the integration order, q is the MA order
   - Estimation- Estimate an ARIMA model using values of p, d, & q
   - Diagnosing checking- checking the residual of estimated model

- Forecasting- Produce out of sample forecasts or set aside last few data points for in-sample forecasting.

This is some example of mathematical model ARIMA model

ARIMA (2,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + b_1 \varepsilon_{t-1}$

ARIMA (3,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + b_1 \varepsilon_{t-1}$

ARIMA (1,1,0) $\Delta y_t = a_1 \Delta y_{t-1} + \varepsilon_t$, where $\Delta y_t = y_t - y_{t-1}$

ARIMA (2,1,0) $\Delta yt = a_1 \Delta y_{t-1} + a_2 \Delta y_{t-2} + \varepsilon_t$ where $\Delta yt = yt - yt\text{-}1$

2. LSTM model

Long Short-Term Memory networks usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn! All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

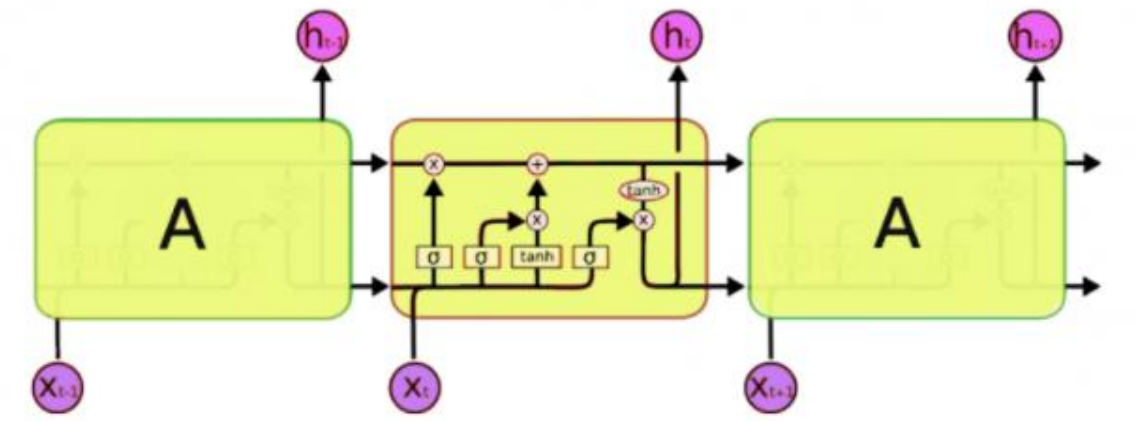


Figure 1 Source: Google

LSTM has a chain structure that contains four neural networks and different memory blocks called cells. Information is retained by the cells and the memory manipulations are done by the **gates.** There are three gates.

1) Forget gate- The information that no longer useful in the cell state is removed with the forget gate. This gate takes in two inputs; h_t-1 and x_t. h_t-1 is the hidden state from the previous cell or the output of the previous cell and x_t is the input at that particular time step. The sigmoid function outputs a vector, with values ranging from 0 to 1, corresponding to each number in the cell state.
2) Input Gate- Addition of useful information to the cell state is done by input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered similar to the forget gate. It creates a vector containing all possible values that can be added (as perceived from h_t-1 and x_t) to the cell state. This is done using the **tanh** function, which outputs values from -1 to +1.

3) Output Gate- The task of extracting useful information from the current cell state to be presented as an output is done by output gate. First, a vector is generated by applying tanh function on the cell. Then, the information is regulated using the sigmoid function and filter the values to be remembered.

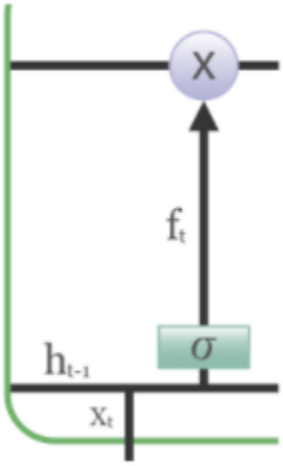


Figure 2 Forget Gate

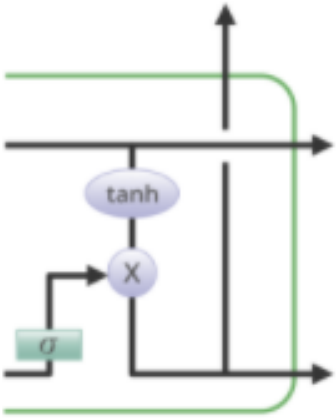


Figure 3 Output Gate

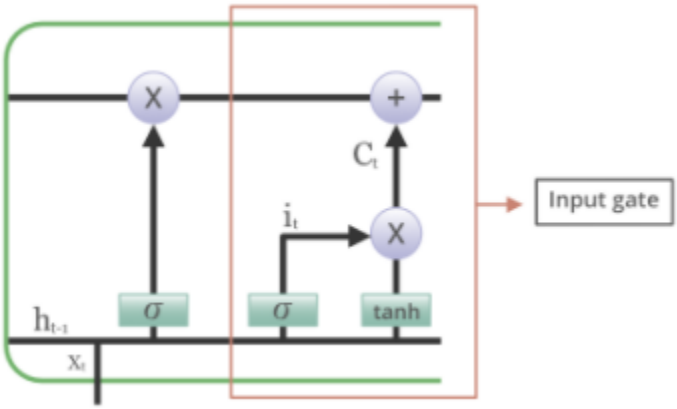


Figure 3 Input Gate

3. Exponential Smoothing Model

The exponential smoothing model will then forecast the future demand as its last estimation of the level. A simple exponential smoothing is one of the simplest ways to forecast a time series. The underlying idea of an exponential smoothing model is that, at each period, the model will learn a bit from the most recent demand observation and remember a bit of the last forecast it did. The previous forecast includes everything the model learned so far based on demand history. The smoothing parameter (or learning rate) **alpha** will determine how much importance is given to the most recent demand observation. Mathematically,

$$f_t = \alpha d_{t-1} + (1-\alpha) f_{t-1}$$

$$0 < \alpha \leq 1$$

If we do a bit of algebra, we obtain the following formula

$$f_t = \alpha d_{t-1} + \alpha(1-\alpha) d_{t-2} + (1-\alpha)^2 f_{t-2}$$

Here you can see that for each further demand observation is reduced by a factor (1-alpha). This is why we call this method exponential smoothing. Here I have used single exponential smoothing model

Results and Analysis

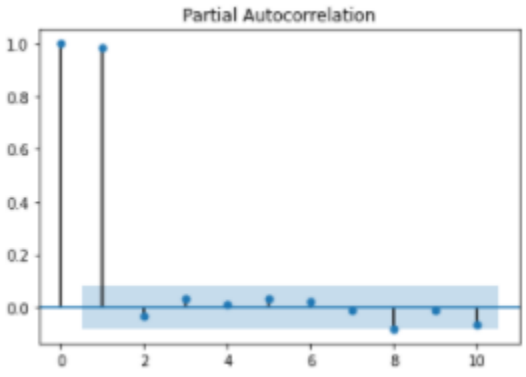
1) Statistical analysis like ACF, PACF, ADF Test and others

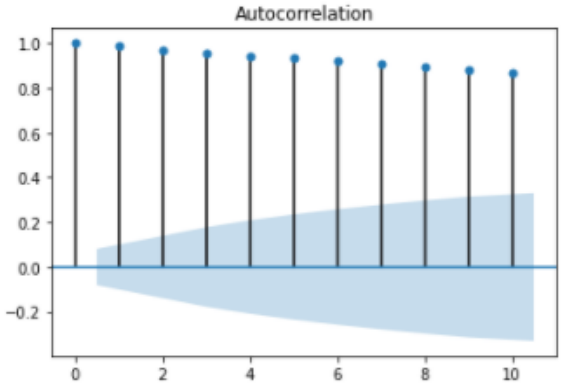
For Apple dataset

```
Results of Dickey-Fuller Test for  Close

Test Statistic                  -2.170605
p-value                          0.217087
#Lags Used                       0.000000
Number of Observations Used    588.000000
Critical Value (1%)             -3.441520
Critical Value (5%)             -2.866468
Critical Value (10%)            -2.569394
dtype: float64
```

```
Results of KPSS Test for  Close
Test Statistic            0.655743
p-value                   0.017569
Lags Used                19.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```





ADF test-
The p value obtained is greater than significance level of 0.05 and test statistics is higher than any of the critical value. So, we can't reject the null hypothesis so the time series is non stationary.

KPSS test-
The p value is significant less than 0.05 hence we can reject the null hypothesis so series is non stationary

ACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

PACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

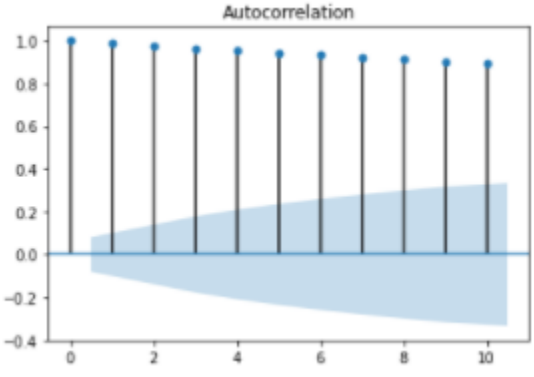
For IBM dataset

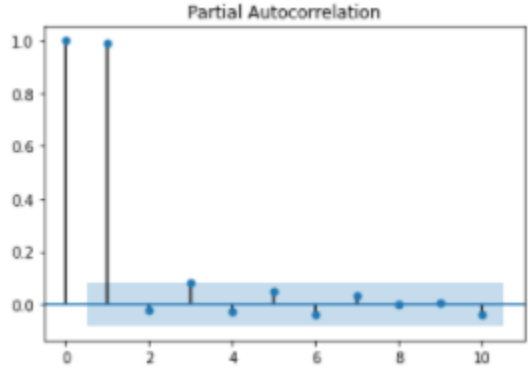
Results of Dickey-Fuller Test for Close

Test Statistic	-2.279273
p-value	0.178740
#Lags Used	0.000000
Number of Observations Used	588.000000
Critical Value (1%)	-3.441520
Critical Value (5%)	-2.866468
Critical Value (10%)	-2.569394

Results of KPSS Test for Close

Test Statistic	1.268862
p-value	0.010000
Lags Used	19.000000
Critical Value (10%)	0.347000
Critical Value (5%)	0.463000
Critical Value (2.5%)	0.574000
Critical Value (1%)	0.739000





ADF test-
The p value obtained is greater than significance level of 0.05 and test statistics is higher than any of the critical value. So, we can't reject the null hypothesis so the time series is non stationary.

KPSS test-
The p value is significant less than 0.05 hence we can reject the null hypothesis so series is non stationary

ACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

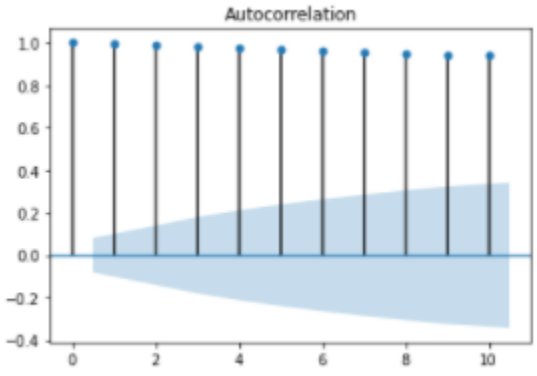
PACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

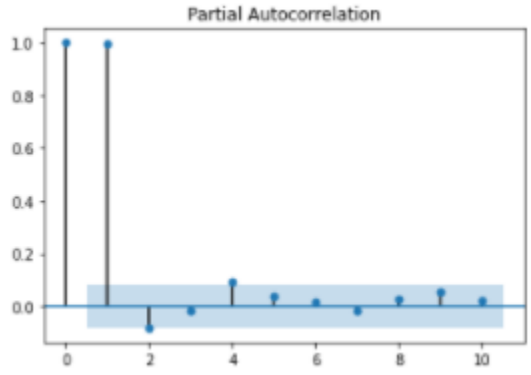
For Google dataset

```
Results of Dickey-Fuller Test for  Close

Test Statistic                  -1.092577
p-value                          0.718023
#Lags Used                       3.000000
Number of Observations Used    585.000000
Critical Value (1%)             -3.441578
Critical Value (5%)             -2.866493
Critical Value (10%)            -2.569408
```

```
Results of KPSS Test for  Close
Test Statistic            2.648792
p-value                   0.010000
Lags Used                19.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```





ADF test-
The p value obtained is greater than significance level of 0.05 and test statistics is higher than any of the critical value. So, we can't reject the null hypothesis so the time series is non stationary.

KPSS test-
The p value is significant less than 0.05 hence we can reject the null hypothesis so series is non stationary

ACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

PACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

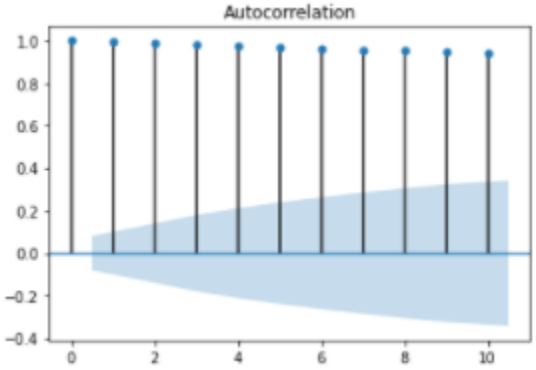
For Facebook dataset

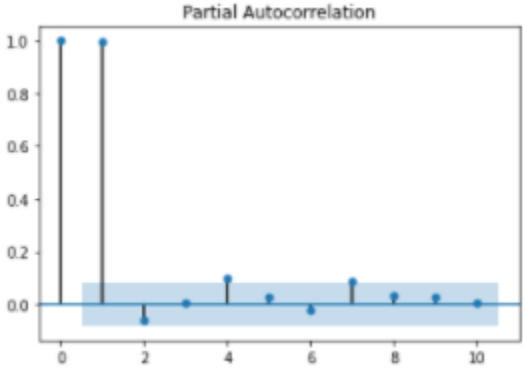
Results of Dickey-Fuller Test for Close

Test Statistic	-0.966644
p-value	0.765235
#Lags Used	6.000000
Number of Observations Used	582.000000
Critical Value (1%)	-3.441636
Critical Value (5%)	-2.866519
Critical Value (10%)	-2.569422

Results of KPSS Test for Close

Test Statistic	2.941636
p-value	0.010000
Lags Used	19.000000
Critical Value (10%)	0.347000
Critical Value (5%)	0.463000
Critical Value (2.5%)	0.574000
Critical Value (1%)	0.739000





ADF test-
The p value obtained is greater than significance level of 0.05 and test statistics is higher than any of the critical value. So, we can't reject the null hypothesis so the time series is non stationary.

KPSS test-
The p value is significant less than 0.05 hence we can reject the null hypothesis so series is non stationary

ACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

PACF test-
As more than 5% of the plot is outside the shaded region, the data is non stationary.

2) Plots of data analysis design and comparison

For Apple dataset

Figure: Exponential

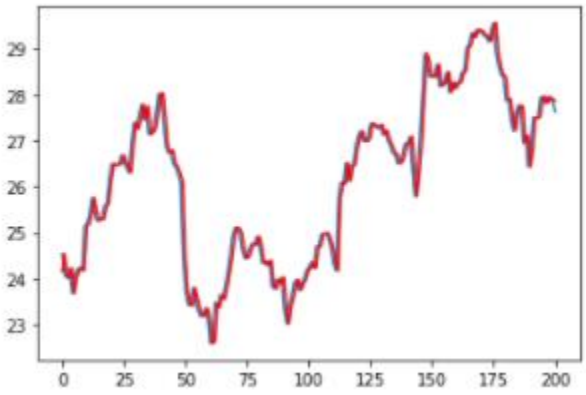


Figure: ARIMA

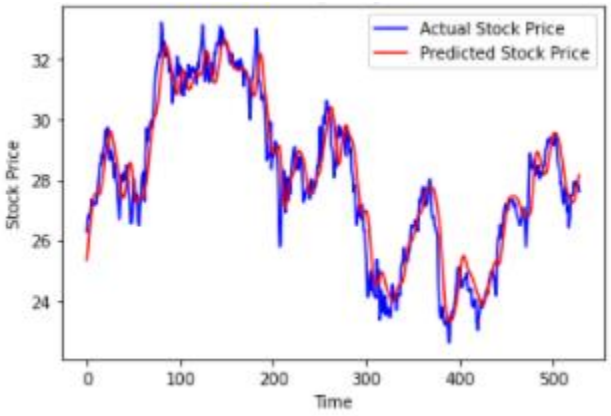


Figure: LSTM

MSE value for Exponential Model is 0.14894212244684166
MSE value for ARIMA Model is 0.12134588182661789
MSE value for LSTM Model is 0.6218421248657574

As MSE value is minimum for ARIMA Model. So, for this dataset ARIMA (1, 0,1) model can be used.

For IBM dataset

Figure Exponential

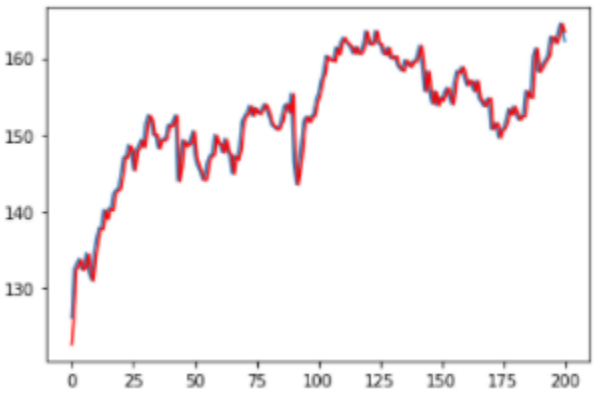


Figure ARIMA

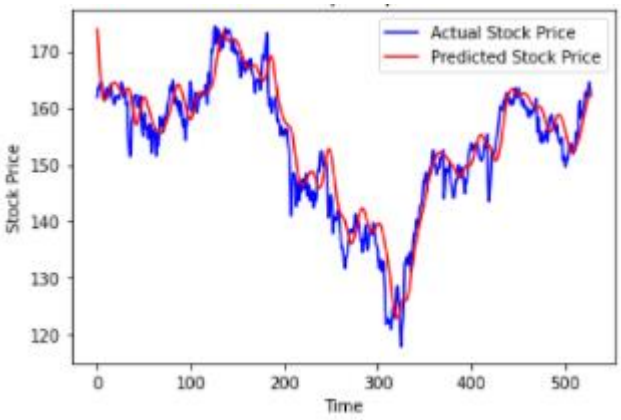


Figure LSTM

MSE value for Exponential Model is 3.4172334666694675
MSE value for ARIMA Model is 3.1982815526323543
MSE value for LSTM Model is 15.992904502728457

As MSE value is minimum for ARIMA Model. So, for this dataset ARIMA (1, 0,1) model can be used.

For Google dataset

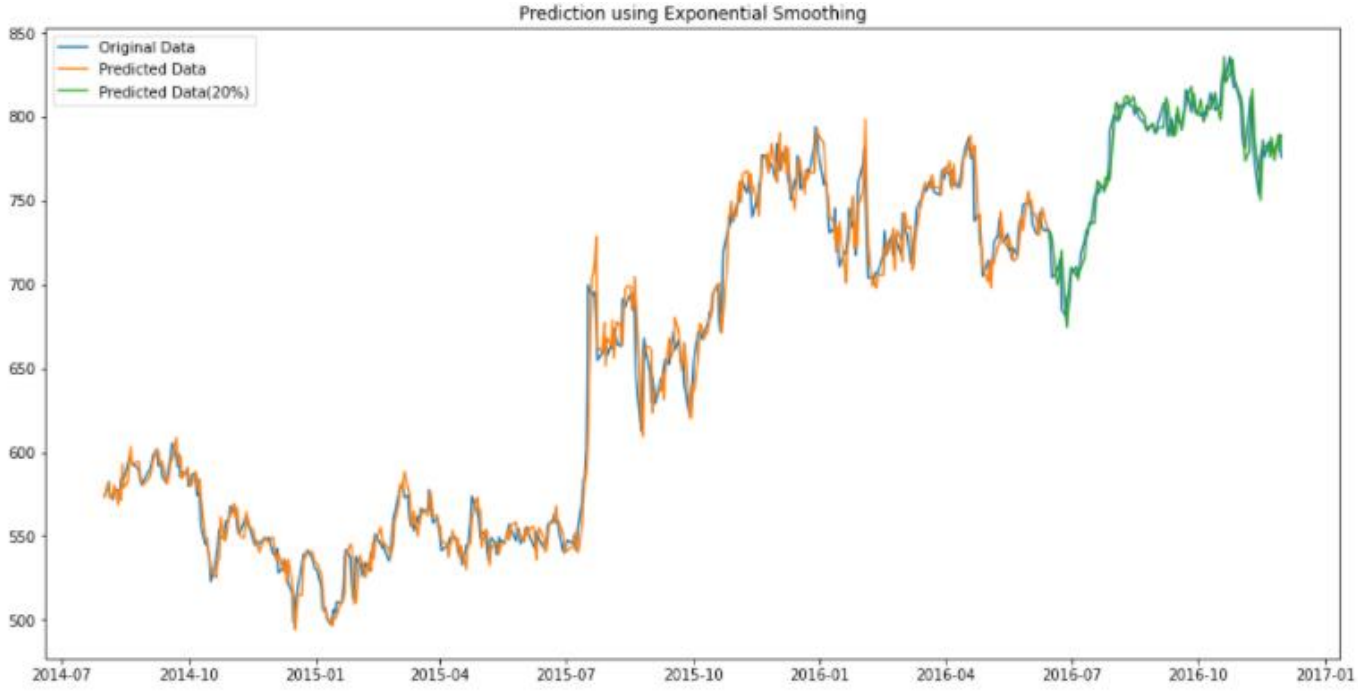


Figure Exponential

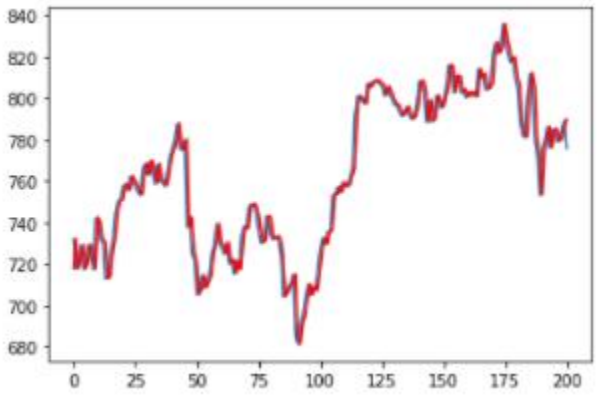


Figure ARIMA

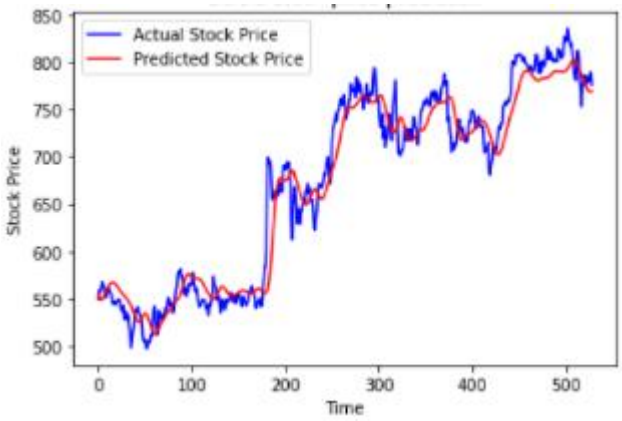


Figure LSTM

MSE value for Exponential Model is 91.8678384539808
MSE value for ARIMA Model is 76.00429273259594
MSE value for LSTM Model is 563.0632418355867

As MSE value is minimum for ARIMA Model. So, for this dataset ARIMA (1, 0,1) model can be used.

For Facebook dataset

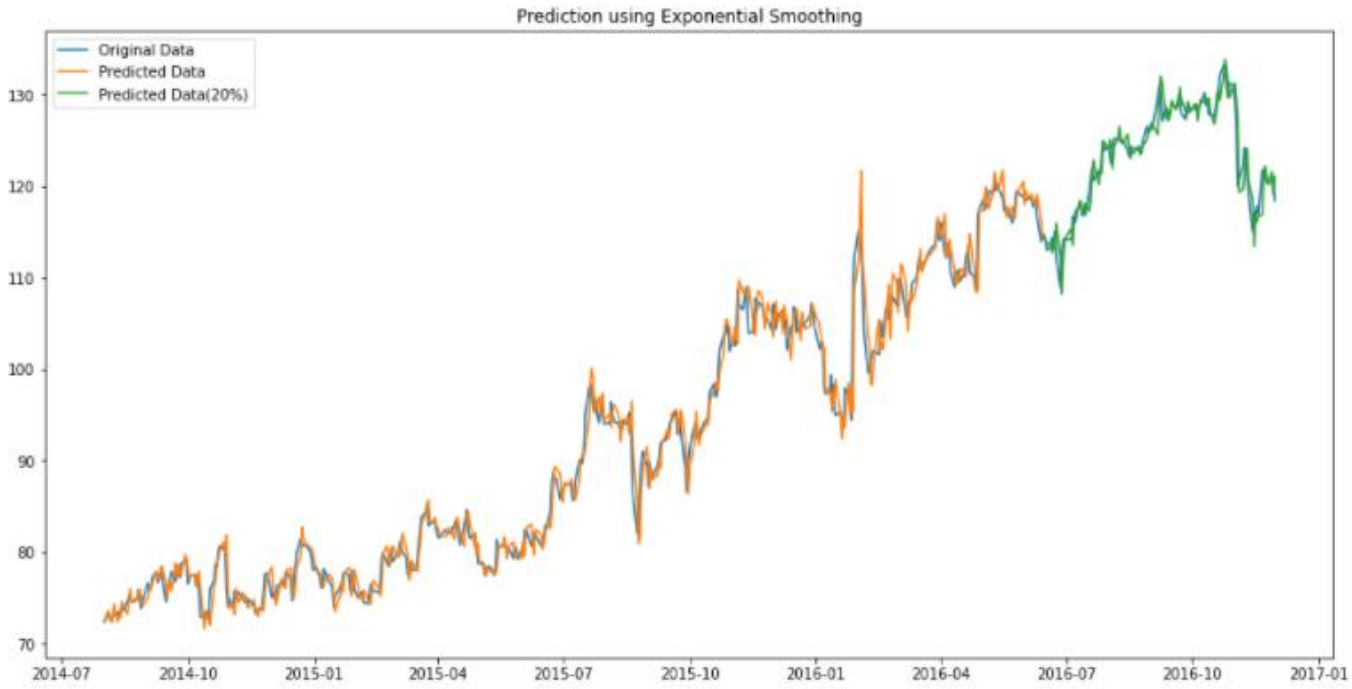


Figure Exponential

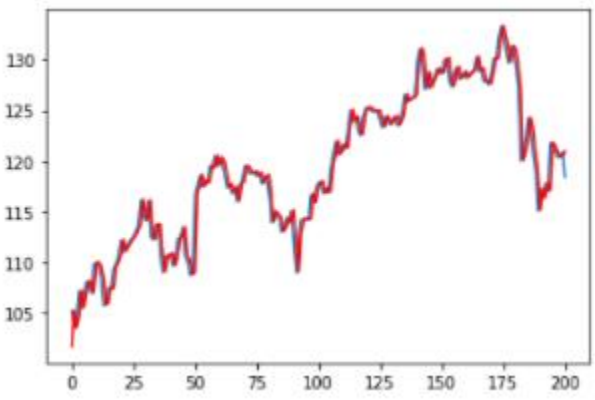


Figure ARIMA

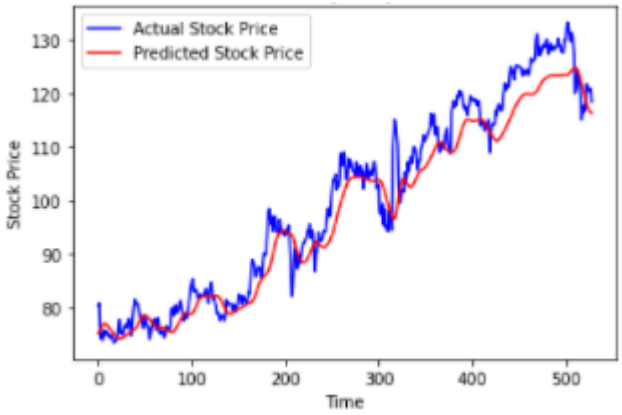


Figure LSTM

MSE value for Exponential Model is 3.126790071926512
MSE value for ARIMA Model is 2.4418359598771464
MSE value for LSTM Model is 20.561681311768684

As MSE value is minimum for ARIMA Model. So, for this dataset ARIMA (1, 0,1) model can be used.

3) implementation files (python, MATLAB, R)

Implemented Jupiter notebook file for all dataset (Apple, IBM, Google, Facebook) is included with this report.

Conclusion

1) On statistical analysis, all stock market dataset that I used show non stationary time series.
2) After applying different type of stock prediction model, we find that **ARIMA (1, 0, 1)** model shows minimum Mean squared error MSE, so it can be used to predict future stock.

Although this is still not a best model exist, there are many advance stock prediction models are available and also here we have considered only the previous stock data but analyzing the market and taking all the effect during the building of any stock prediction model will give the best results.