Stock Prediction of Facebook

By
Karishma
Ravi Teja Buddabathuni
Keshav Anand Botlaguduru
Sruthi Vanam

INTRODUCTION:

Facebook is an American online social media and social networking service based in Menlo park California. It was founded in 2004 by Mark Zuckerberg along with his fellow Harvard students. It is having multiple

child companies like WhatsApp, Instagram. Facebook is leader in social networking. They wanted to deploy servers all over the world and for that reason they needed investors. So, in the year of 2012 Facebook entered stock market where they started their Initial Public Offering and began to sell their stock after 3 months from US$ 31.71 reaching and original peak market capitalization of US$ 104 billion and then they became the biggest social networking sites which now having a stock price of US$ 150.25. Now Facebook is having more that 1.86 million active users all over the world.

OBJECTIVE:

Here we are going to predict stock price of Facebook based on the historical data using Time series model. We will be comparing various model and then we will be checking goodness of our selected model and based on that we will be performing stock forecasting to help the investors to take the timely decision while choosing this stock.

MEET THE DATA:

For prediction, we took data from 2013 to December 2016.Our dataset had 1008 observations and 7 variables. We have imported the dataset in R using quantmod from yahoo finance.

	Date	Open	High	Low	Close	Volume	Adjusted Close
1	2013-01-02	27.44	28.18	27.42	28.00	69846400	28.00
2	2013-01-03	27.88	28.47	27.59	27.77	63140600	27.77
3	2013-01-04	28.01	28.93	27.83	28.76	72715400	28.76
4	2013-01-07	28.69	29.79	28.65	29.42	83781800	29.42
5	2013-01-08	29.51	29.60	28.86	29.06	45871300	29.06

Date: Date of observation

Open: Opening price for a date.

High: High price of a day.

Low: Low price of a day.

Volume: Total number of stocks sold for a date.

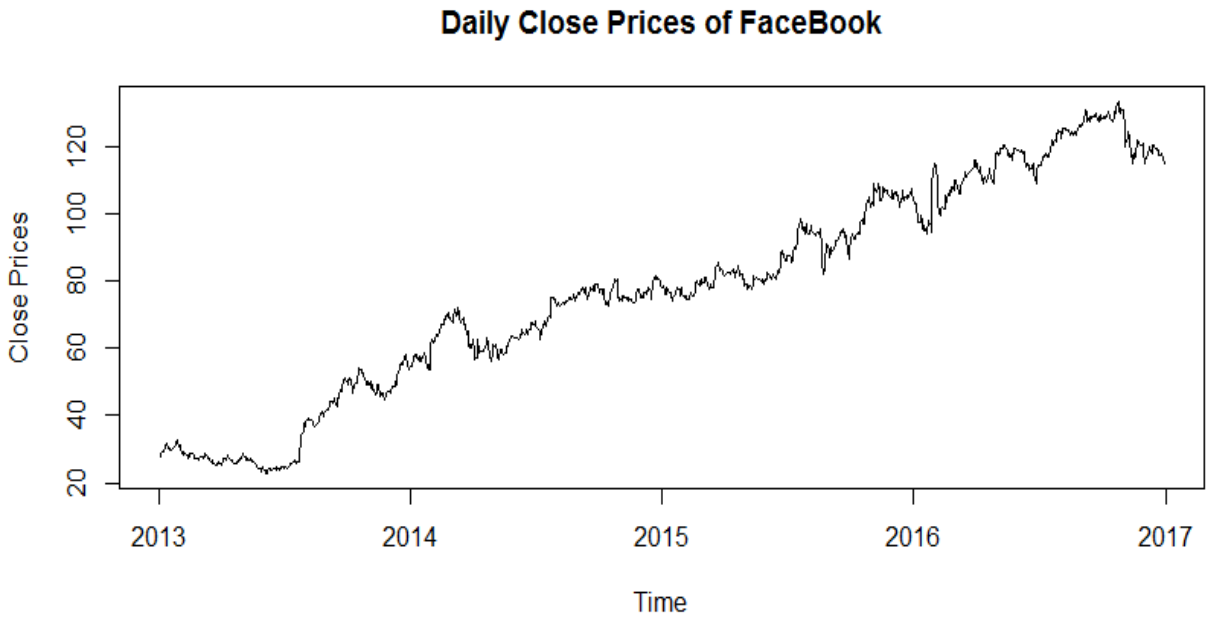
Adjusted Close: After the closing of stock, if there are corporate changes like stock splits, right offerings then it is captured in adjusted close.

CHALLENGES FACED:

Since we took data from yahoo using quantmod there were missing values as stock market does not run on Saturday and Sunday. So, to manage missing values we created continuous time series object by exporting the imported time series object to an Excel and loading that excel as continuous time series object. To create time series model, we had to select model and we had selected close price as variable instead of adjusted close because adjusted close depends on corporate actions which we cannot predict hence we selected close price as variable.

EXPLORATORY DATA ANALYSIS:

To build a time series the primary condition is that the data should be stationary. Here we plotted time series plot for close price.



So here we can see that the data is not stationary and it is having upward trend. Also, there is no seasonality because there no repetition of trend monthly, quarterly or yearly.

To further check about the stationarity we have performed *adf test*.

```
> adf.test(close_stock, alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  close_stock
Dickey-Fuller = -2.9316, Lag order = 10, p-value = 0.1839
alternative hypothesis: stationary
```

Null Hypothesis: It is not Stationary

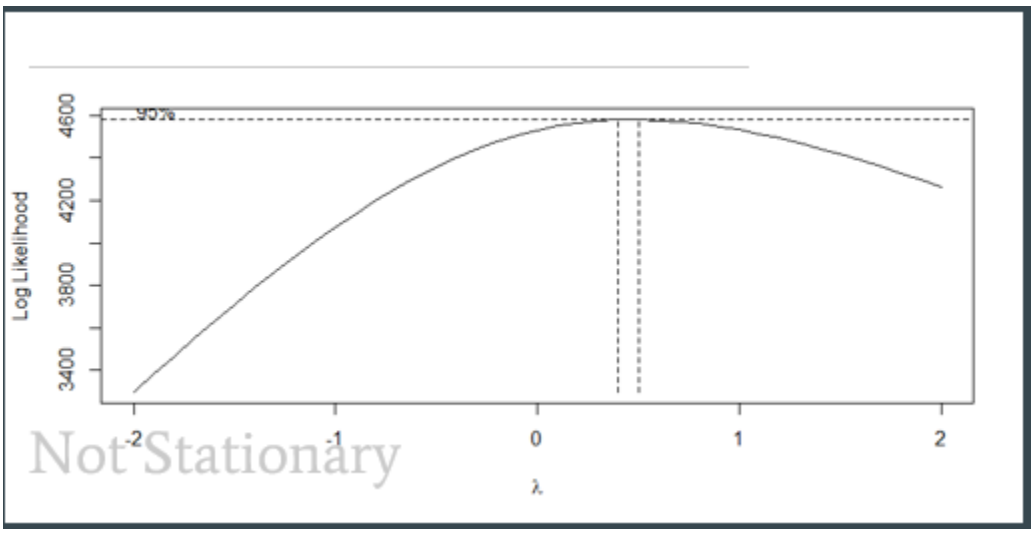
Alternative Hypothesis: Stationary

From the above output, we can see that ***p > α*** hence we failed to reject our Null Hypothesis.

That means data is not stationary.

Box Cox Transformation:

To make data stationary we need to transformation. With the help of box cox curve, we can know about the transformation required to make data stationary.



From here we got the value of transformation to be 0.5.

In the later steps we transformed the data using a power transformation of 0.5 in order to smoothen the data.

TRANSFORMING AND DIFFERENCING:

To make the data stationary, the data of close stock price is transformed to square root, and took a lag differencing of 1. After that the differenced data is stationary.

To check the stationarity of the transformed data Augmented Dickey-Fuller Test has been performed and the results were as below,

```
> adf.test(sd1, alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  sd1
Dickey-Fuller = -10.104, Lag order = 10, p-value = 0.01
alternative hypothesis: stationary
```

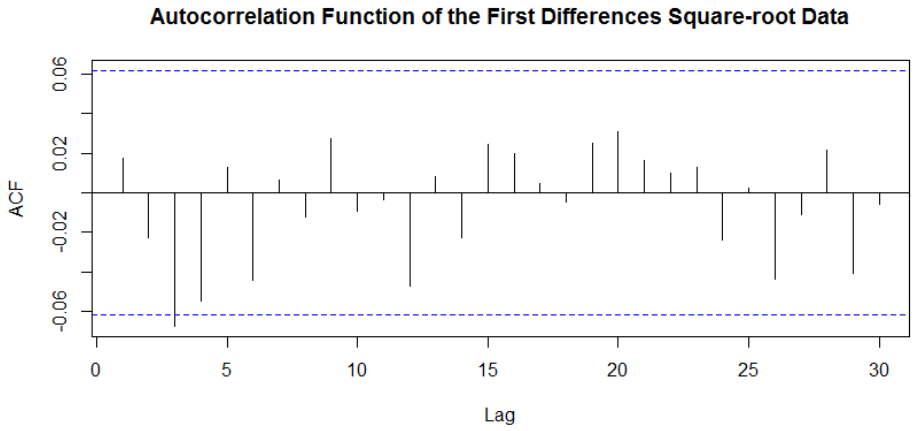
From the above result, we can say that the transformed data is stationary by considering the p-value. Here the p-value is less than alpha (0.05). From this we can directly reject the null hypothesis which is the data is not stationary. That means we are accepting the alternate hypothesis which means the data is stationary.

MODEL IDENTIFICATION:

After stationarizing the time series by transferring and differencing the next step is to fit a Time series ARMA model. For determining the AR and MA terms of the plots for ACF and PACF were plotted.

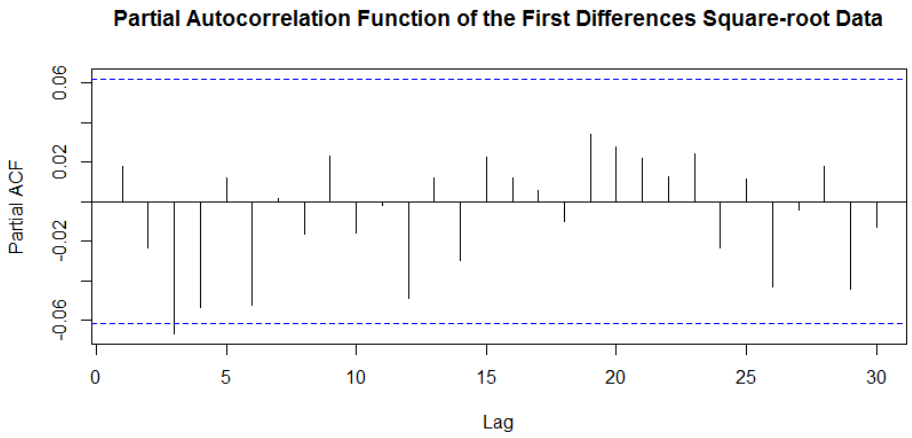
ACF plot:

From the below ACF plot we can see that the data is almost independent and it doesn't show any particular auto correlation across the time. So in this case the next step is to try to interpret the PACF plot.



PACF plot:

From the below PACF plot we tried to see the partial auto correlation of the data. Even from this plot the data is looking like independent and contains white noise.



MODEL SELECTION:

Before Proceeding to the random walk model, a preliminary check for model has been done by using the auto.arima function and from that a suggestion of ARIMA (1,1,4) as below

```
> auto.arima(close_stock)
Series: close_stock
ARIMA(1,1,4) with drift

Coefficients:
          ar1     ma1      ma2      ma3      ma4   drift
      -0.9408  0.9643  -0.0089  -0.1316  -0.1321  0.0867
s.e.   0.0378  0.0487   0.0447   0.0460   0.0342  0.0406

sigma^2 estimated as 2.194:  log likelihood=-1821.52
AIC=3657.04   AICc=3657.15   BIC=3691.44
>
```

From the above suggested model, AIC value is 3657.04

From the details of the ACF and PACF plots a Random walk model is built by using the ARIMA(0,1,0)

Which has an AIC value of 3669.69

The equation for the model is as below:

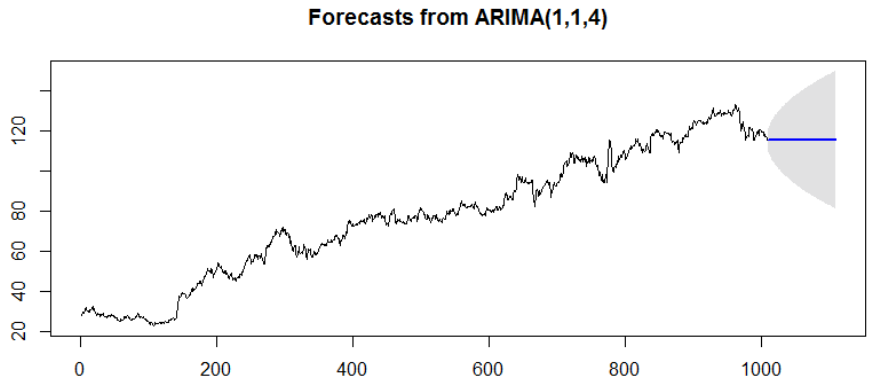
$$Yt = yt\text{-}1 + e(t)$$

```
> model2 = Arima(close_stock,order=c(0,1,0))
> summary(model2)
Series: close_stock
ARIMA(0,1,0)

sigma^2 estimated as 2.025:  log likelihood=-2193.33
AIC=4388.66   AICc=4388.66   BIC=4393.78

Training set error measures:
                     ME      RMSE       MAE         MPE      MAPE      MASE        ACF1
Training set -0.08499299 1.422382 0.9754267 -0.1362613 1.589238 0.9993117 0.03063314
>
```

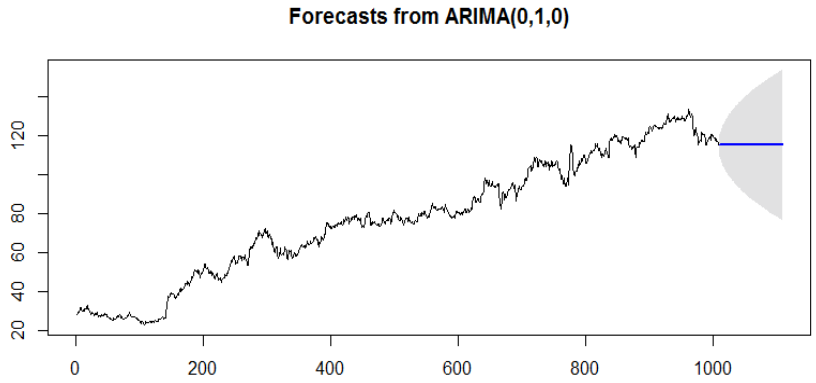
Although by verifying the AIC values the Model 1 which is ARIMA(1,1,4) is best we tried to plot the forecast of these models and wanted to compare with the Actual stock prices.

The below steps provide a detailed summary about the forecast of these both models.

By using the R forecast function a forecast has been drawn for the model 1 which shown as below and having a same value for all the upcoming values. In this it is not following any trend and this forecast is done with a 99% confidence interval



By using the R forecast function a forecast has been drawn for the model 2 as well which shown as below and having a same value for all the upcoming values. In this it is not following any trend and this forecast is done with a 99% confidence interval.



Forecast Comparison with Actual values:

Forecasted values from Model 1

```
> forecast(object = model1, h = 100, level = 99)
     Point Forecast     Lo 99    Hi 99
1009       115.1329 111.31135 118.9545
1010       115.2830 109.79950 120.7666
1011       115.3582 108.66582 122.0506
1012       115.4434 107.90584 122.9809
1013       115.3631 107.12041 123.6057
1014       115.4388 106.49922 124.3783
1015       115.3674 105.82328 124.9115
1016       115.4347 105.28490 125.5845
1017       115.3713 104.68318 126.0594
1018       115.4310 104.20066 126.6614
1019       115.3747 103.65392 127.0955
1020       115.4278 103.21199 127.6436
1021       115.3777 102.70829 128.0472
1022       115.4249 102.29740 128.5525
1023       115.3805 101.82867 128.9322
```

Forecasted values form Model 2

```
> forecast(object = model2, h = 100, level = 99)
     Point Forecast     Lo 99    Hi 99
1009         115.05 111.19903 118.9010
1010         115.05 109.60390 120.4961
1011         115.05 108.37992 121.7201
1012         115.05 107.34805 122.7520
1013         115.05 106.43896 123.6610
1014         115.05 105.61708 124.4829
1015         115.05 104.86128 125.2387
1016         115.05 104.15780 125.9422
1017         115.05 103.49707 126.6029
1018         115.05 102.87215 127.2279
1019         115.05 102.27776 127.8222
1020         115.05 101.70983 128.3902
1021         115.05 101.16511 128.9349
1022         115.05 100.64097 129.4590
1023         115.05 100.13524 129.9648
```

Actual values

```
> FB
           FB.Close
2017-01-03   116.86
2017-01-04   118.69
2017-01-05   120.67
2017-01-06   123.41
2017-01-09   124.90
2017-01-10   124.35
2017-01-11   126.09
2017-01-12   126.62
2017-01-13   128.34
2017-01-17   127.87
2017-01-18   127.92
2017-01-19   127.55
2017-01-20   127.04
2017-01-23   128.93
2017-01-24   129.37
2017-01-25   131.48
2017-01-26   132.78
```

Actual Values of Close stock price those are compared to the forecasted values with both the models

Here from the both models the values which are forecasted are not related to the actual values, so to overcome with this a random walk model with drift is included in the model selection.

RANDOM WALK

Random walk is a statistical theory which is used to understand a variable which follows no particular pattern or trend. The random walk theory states that the movement of the value of variable, say y(t) is random and hence non predictable. The value of y(t) is only is hence y(t-1) plus an error term which gets added as time goes on.

Hence,

$y(t) = y(t-1) + e$

Where e is the error term, t and (t-1) are two points in time scale.

Since random walk states that the movement of variable “y” across time “t” is completely random, the variance and covariance increase as a function of time t.

Random walk is used in multiple real world models. Particle physicists use random walk to simulate the movement of atoms in gaseous state in order to develop physical boundaries and energy settings for gas chambers. By doing so, they can typically identify the limits to which the experiment can vary and can plan their contingencies. In stock market as well, where change in prices don’t seem to follow any deterministic pattern, random walk is the best way to take a guess. By using random walk models, we can equate the historical data and understand the extent to which error terms vary. Based on this, we can understand

the extremities within which the variable will fluctuate. Random walk does not predict the variable value at any particular point in time but only predict the extent to which the variable might vary. Hence random walk simulations give different results when executed at different times.

Future price of a stock is always speculated and hence billions of dollars of trading happens across various stock markets around the world every day. Historically there have been statistical ways in which stock prices have been speculated. These methods include:

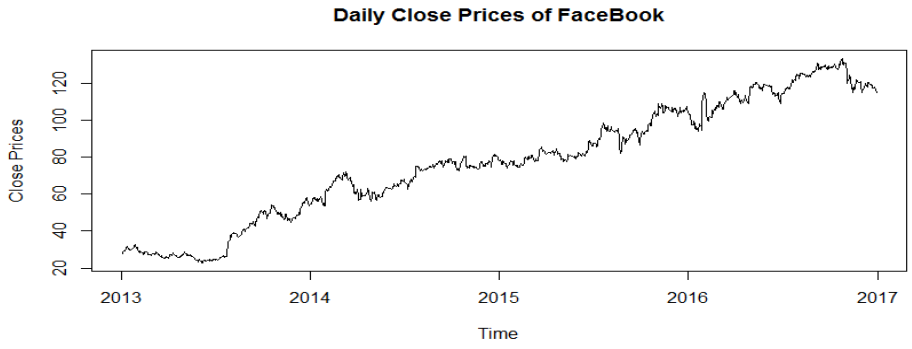
1) Charting
2) Calculation of intrinsic value.

In charting, the principle used is that history repeats itself. Hence the seasonlaity in data is used as a principle concept in forecasting a stock value. Also, patterns which coincide with external conditions like political circumstances are taken into consideration. In the second method, value of a stock is calculated using complex algorithms and is compared with the current stock value. If the current stock is undervalued, then investing into the stock is done and on the other hand if the stock is overvalued then short stocking is done. Short stocking is a method in which shares can be sold first and bought later.

Apart from these techniques, many non-statistical methods are also used to predict stock value fluctuation. Staying in touch with the perception that a company gives out and understanding market factors like demand and supply lead to speculation of stock prices. This is exactly where random walk is different from other statistical processes to predict stock value. Random walk doesn't have any procedure of forecasting the stock price but only uses the historical data to understand how much the error terms, or in this case the change in stock price per day, vary. Based on this, a certain confidence interval is created within which the stock price is speculated to fluctuate. Based on this interval, stocks are classified as low risk, high risk and volatile stocks and investors use this technique to classify various stocks within a cluster of multiple stocks which is marketed as a "portfolio"

Hence, we implemented the random walk method in our stock value prediction.

In the data that is taken, by looking at the historical data it is clear that there is a gradual upward trend in the value of the stock. Hence this trend is taken into consideration when predicting the future value of the stock. The figure below provides a clear illustration of this.

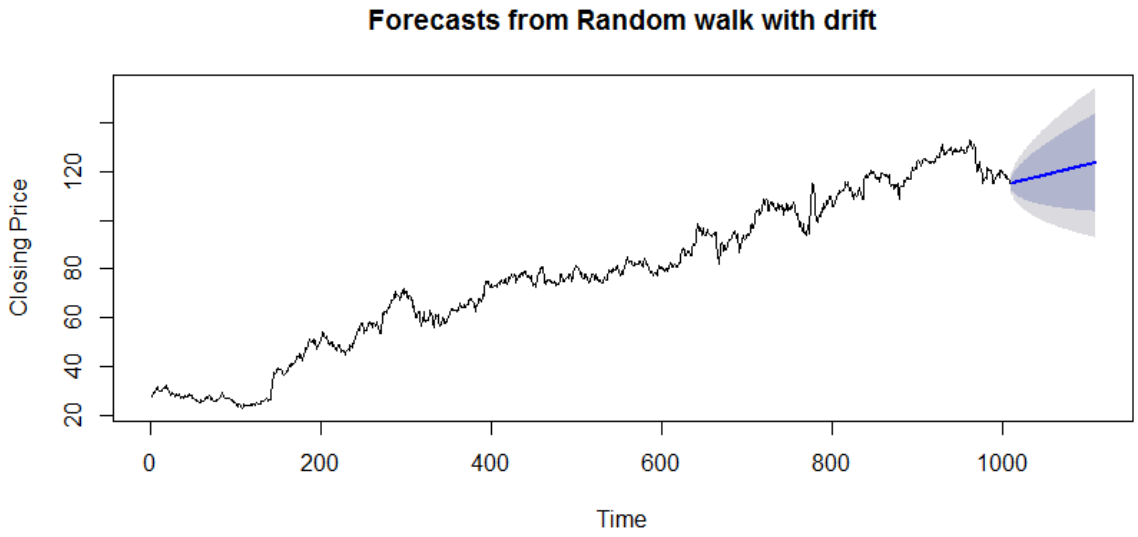


To factor this trend into consideration, drift is taken into the random walk model. In a random walk model which comprises of drift, the value y(t) is:

Y(t) = D + Y(t-1) + e

Where D is the drift. Hence drift quantizes the trend as a function of time. Every time the value of Y changes, the drift adds to it, which preserves the trend in the future prediction. After calculating the drift for the random walk with drift model, we infer that the value of drift is 0.0864. This means that with every additional time "T", the value of Y increases by 0.0864.

```
> summary(rwalkfore)

Forecast method: Random walk with drift

Model Information:
Call: rwf(y = close_stock, h = 100, drift = T, level = 99)

Drift: 0.0864  (se 0.0471)
Residual sd: 1.4933

Error measures:
                      ME     RMSE      MAE          MPE     MAPE
Training set 3.520967e-15 1.492542 1.023192 -0.02334237 1.443462
```



$$Y_t = y_{t-1} + e(t) + \theta_o$$

By observing the above picture, we can clearly see an upward trend being factored into the forecast. The forecasted and real values are placed next to each other below.

(paste values)

The model clearly aligns with the real data in the fact that there is an upward trend. Even the Forecasted values were describing the data set properly as we are able to predict in the confidence intervals. Since the original values are in the confidence intervals of the forecasted values, this model is best suited for Facebook Data.

```
Forecasts:
     Point Forecast     Lo 99    Hi 99
1009       115.1364 111.29000 118.9829
1010       115.2229 109.78050 120.6653
1011       115.3093 108.64049 121.9782
1012       115.3958 107.69145 123.1001
1013       115.4822 106.86425 124.1002
1014       115.5687 106.12349 125.0139
1015       115.6551 105.44811 125.8621
1016       115.7416 104.82443 126.6587
1017       115.8280 104.24293 127.4131
1018       115.9145 103.69670 128.1322
1019       116.0009 103.18050 128.8213
1020       116.0873 102.69029 129.4844
1021       116.1738 102.22285 130.1247
1022       116.2602 101.77556 130.7449
1023       116.3467 101.34627 131.3471
```

```
> FB
           FB.Close
2017-01-03   116.86
2017-01-04   118.69
2017-01-05   120.67
2017-01-06   123.41
2017-01-09   124.90
2017-01-10   124.35
2017-01-11   126.09
2017-01-12   126.62
2017-01-13   128.34
2017-01-17   127.87
2017-01-18   127.92
2017-01-19   127.55
2017-01-20   127.04
2017-01-23   128.93
2017-01-24   129.37
2017-01-25   131.48
2017-01-26   132.78
```

RESIDUALS :

Residuals in forecasting is the difference between an observed value and its forecast based on other observations

$$e_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is the prediction of y_i based on all observations except y_i

To make sure that our model is a good forecasting methods, we should have see that the residuals are having following properties:

The residuals should be uncorrelated. if we see any correlation, that means there is some information in the residuals which has to be used for forecasting.

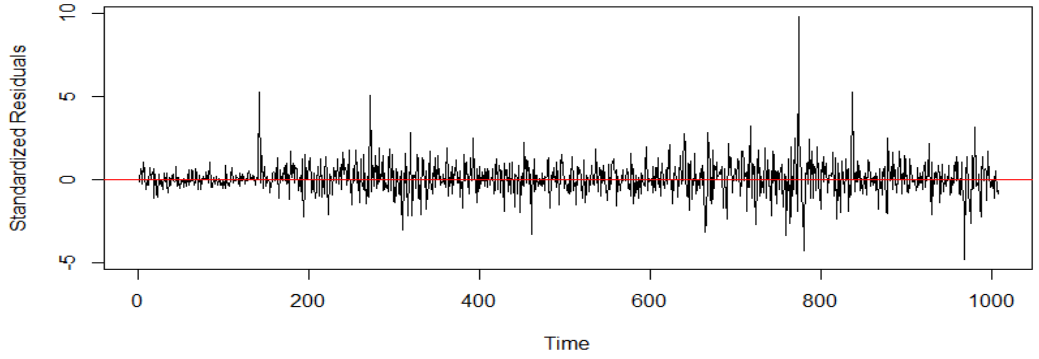
The residuals to have zero mean. if it does not have zero means, it implies that the forecasts are biased.

- It is better to have residuals follow normal distribution.
- It is better if the residuals are having constant variance.

To check the above properties of residuals, different tests like Anderson darling normality test or Shapiro test, qq plots, Ljung-Box test.,

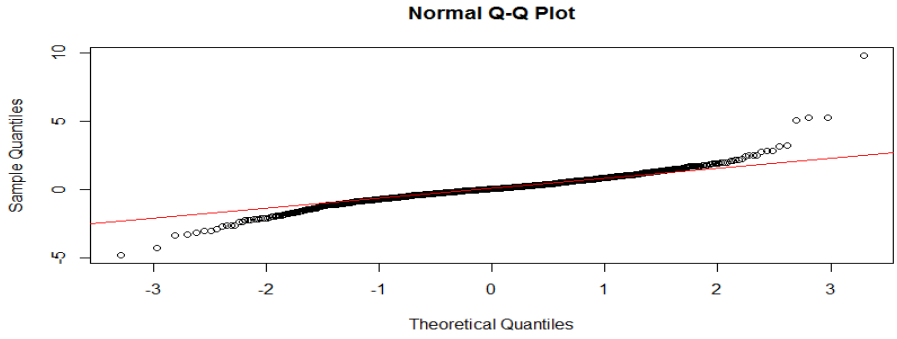
PLOT FOR STANDARDIZED RESIDUALS:

When we plotted the residuals against time, we got the following graph, where we see that the residuals are not following any pattern.



QQ plot for Residuals

Quantiles of the theoretical data and the sample data were plotted against each other onto a straight line as shown below:



Here we see that most of the points falls on the qq line except few outliers which indicates it mostly has a normal distribution, which can be confirmed by performing AD test on it.

Anderson-Darling test:

Ho: residuals are normally distributed

Ha: residuals are not normally distributed.

```
> ad.test(rwalkfore$residuals)

        Anderson-Darling normality test

data:  rwalkfore$residuals
A = 11.89, p-value < 2.2e-16
```

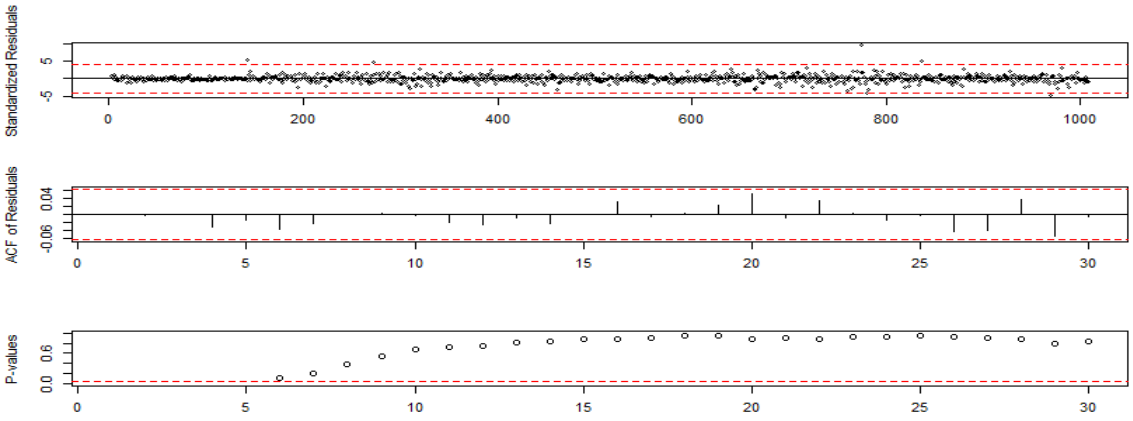
From the above R output of ad test, we see P values is less than α value (0.05) which implies we reject the null hypothesis saying that the residuals are not normally distributed. This property of having normal distributed is a useful property but not necessary as it only helps in making the calculation of our prediction intervals easier. we can try converting these residuals into normal distributed form by applying different transformations like Log etc., on it.

Ljung-Box Test :

Ljung box test is used to test whether there is any auto-correlation in the data and it tests the overall randomness based on the number of lags.

Ho: The data are independently distributed, which means The residuals have no auto correlation

Ha: The data are not independently distributed, which means the residuals have auto correlation



In the above plot obtained from Ljung test:

1. In the first plot, we see that standardized residuals are distributed over the zero mean.
2. In the second plot, we see that ACF values are within the range which indicates there is no auto correlation between the residuals.
3. In the third plot, we see P value is greater than the significant value (0.05) which means we fail to reject the null hypothesis. Therefore, we can say that the residuals are independently distributed and are not correlated.

By summarizing the all above points we can say that residuals of our model are satisfying all the properties and therefore our model is a good model.

CONCLUSION :

- After considering all the factors like AIC values, forecast values etc.., shown above, we can say that Random walk with drift is the better model.
- As we know that the stock prices are volatile, we cannot accurately predict the stock price but this can help in taking any timely decisions.
- Stock prices also depends on many other factors like brand names, brand value, managers, news related to the brand, further analysis can be done using these factors to make a better model. Like we can also perform text mining on news data to increase the accuracy.

Finally, we can say that even though FB stocks are volatile in short term. This can be a good investment in long term.