STOCK PRICE PREDICTION USING TIME SERIES FORECASTING

White paper

OCTOBER 16, 2022

SUBHASHINI NATARAJAN

BUSINESS PROBLEM

Stock trading is an important financial activity for every business and purchasing stocks is an important investment means for individuals. Amazon is one of the promising stocks today and the paper is about predicting Adjusted Closing Rate of Amazon stocks. This helps in understanding the probability of long-term investment returns. The stock forecasting is performed using time series analysis, where the forecasting of ACR can be performed.

BACKGROUND/HISTORY

Stock prediction can be performed for forecasting long term returns or short-term returns. For the project, predicting a long-term stock price for the portfolio, helps us understand the direction in which portfolio is trending. Even though there are a lot of factors that play a role in portfolio performance and accurate predictions of the stocks are not always possible, time series gives us a sense of understanding of the portfolio.

DATA EXPLANATION

Amazon stock data is obtained from yahooFinance package using the Ticker code, “AMZN”.

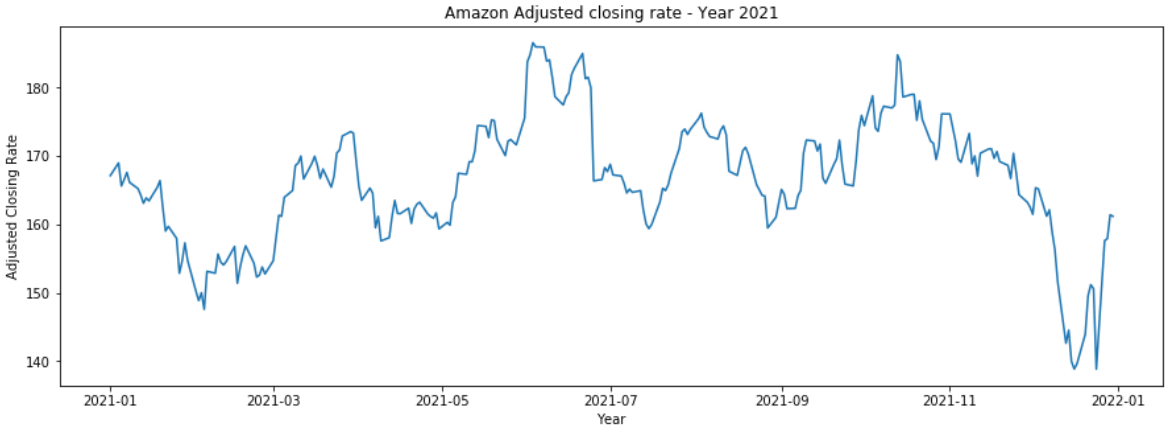
The data set imported includes the following fields-

- Date – Date of business transaction
- Open – Opening Rate
- High – Highest Rate
- Low – Lowest Rate
- Close – Closing Rate
- Volume - Volume of shares

Among the given data, only date field and Closing rate has been chosen for the time series analysis. Starting date for the dataset is 2021/01/01 and End date for the dataset is 2022/05/31.

METHODS & ANALYSIS

As only two features are required for the time series analysis, no significant data preparation activities were required. Dataset was split into train and test, with training data comprising until 2021/12/31 and test data greater than 2021/12/31. Exploratory data analysis was performed with this data set.

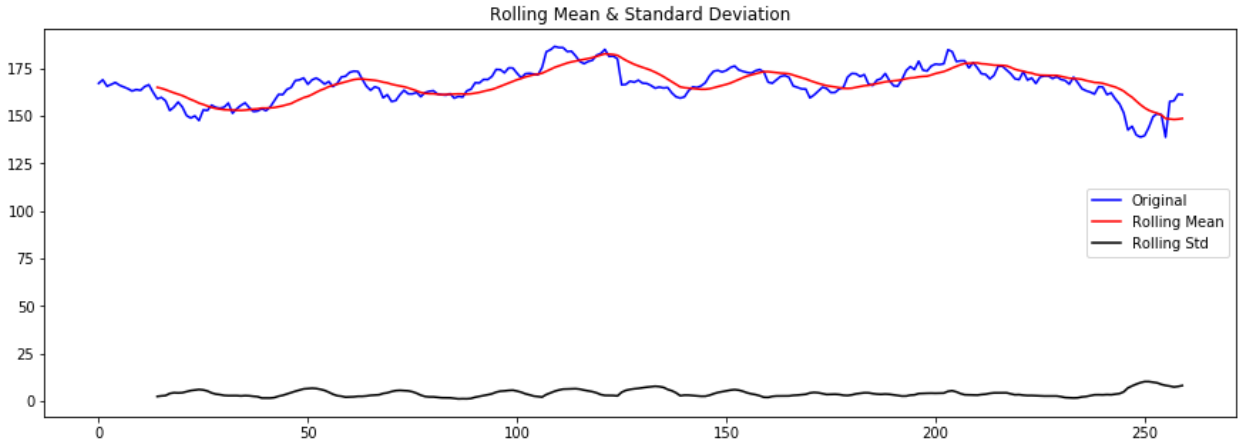


A time series model requires the data to be stationary and without auto correlation. The data is first evaluated for its stationary nature. This is achieved by Dickey Fuller test. This test, in null hypothesis, assumes that the data is nonstationary. The results of the test include test statistic and critical values. If the test statistic value is lower than the critical value, then null hypothesis is not true, and we opt for alternate hypothesis.

For the training data set, the results of Dickey Fuller test are as below -

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.595573
p-value                          0.093916
```

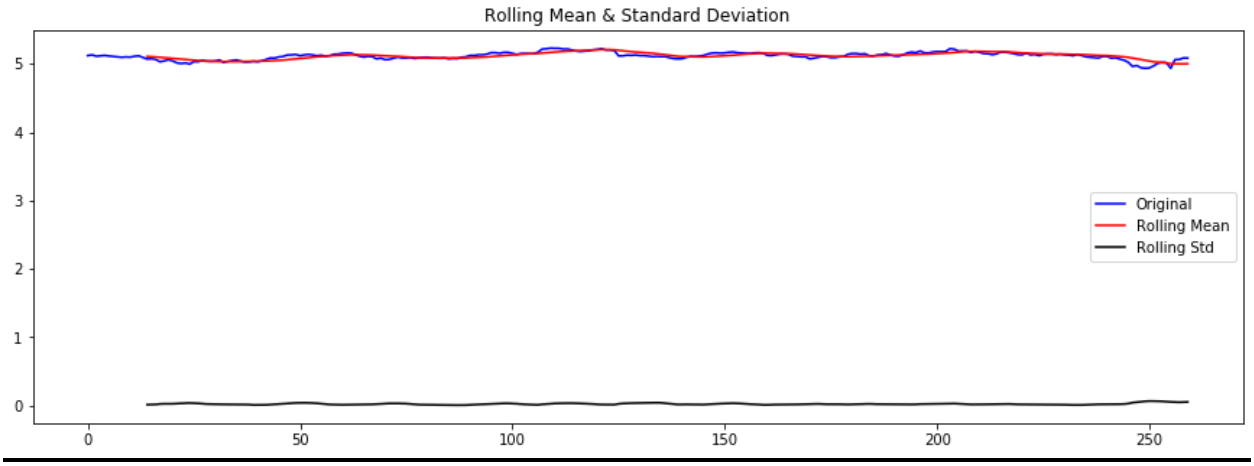
```
#Lags Used                          0.000000
Number of Observations Used       259.000000
Critical Value (1%)                -3.455853
Critical Value (5%)                -2.872765
Critical Value (10%)               -2.572752
dtype: float64
```



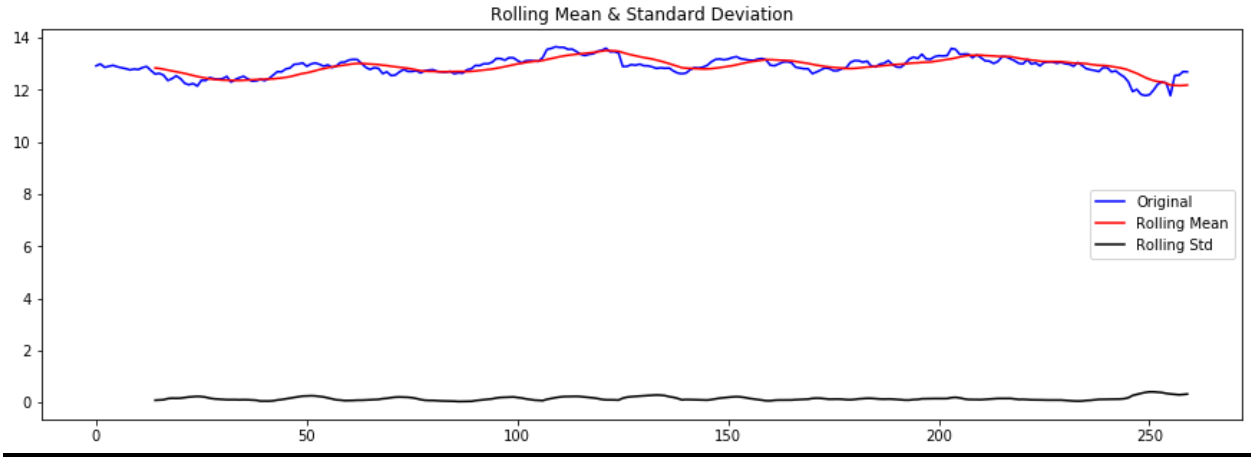
The results show that the data is not stationary. Hence, to make the data stationary, first log transform and then square root transform are applied.

Dickey Fuller Test result after log transformation:

```
Test Statistic                    -2.653978
p-value                            0.082346
#Lags Used                         0.000000
Number of Observations Used      259.000000
Critical Value (1%)               -3.455853
Critical Value (5%)               -2.872765
Critical Value (10%)              -2.572752
dtype: float64
```



Dickey Fuller Test result after Square root transformation:



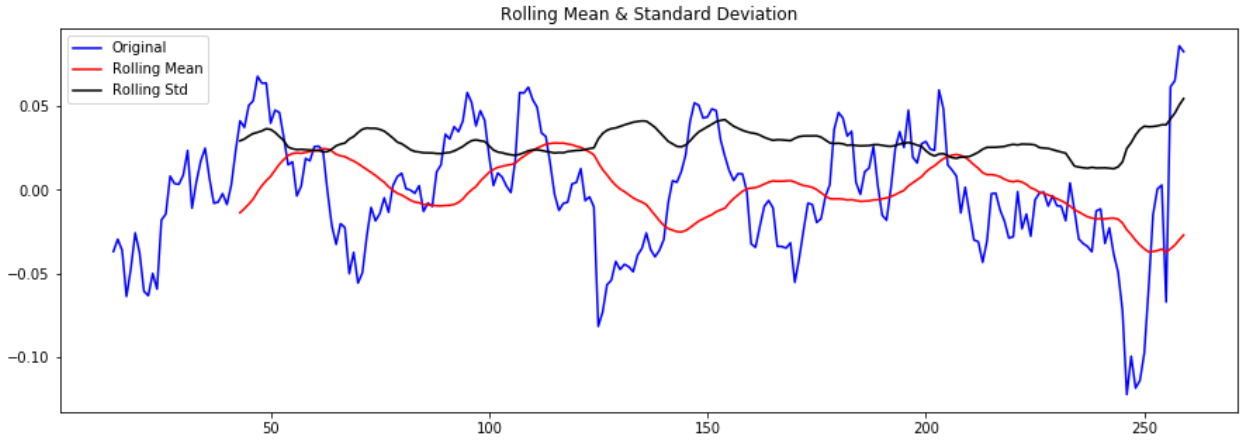
Rolling mean and standard deviation are constant as we see the plots above. However, in the results of this Dickey-Fuller test, test statistics is greater than the critical values, which means that the null hypothesis holds true, and the log transformed object is non-stationary in nature.

The next step in making the data stationary is to estimate the trend and removing it from the original series.

Dickey Fuller test after moving average method:

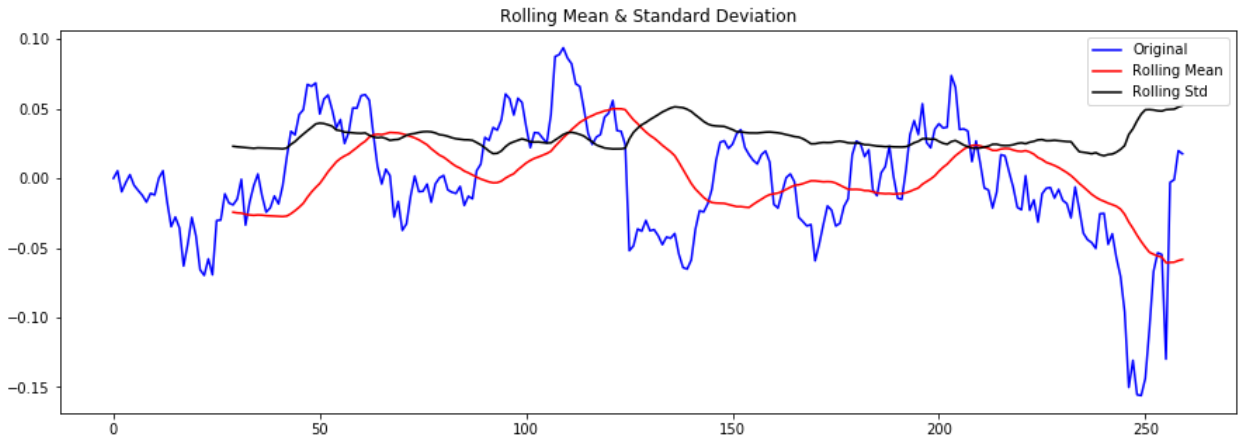
```
Test Statistic                  -3.696958
```

```
p-value                          0.004156
#Lags Used                       0.000000
Number of Observations Used    245.000000
Critical Value (1%)             -3.457326
Critical Value (5%)             -2.873410
Critical Value (10%)            -2.573096
dtype: float64
```



Dickey Fuller test after exponentially weighted moving average method:

```
Test Statistic                  -3.389500
p-value                          0.011322
#Lags Used                       0.000000
Number of Observations Used    259.000000
Critical Value (1%)             -3.455853
Critical Value (5%)             -2.872765
Critical Value (10%)            -2.572752
dtype: float64
```



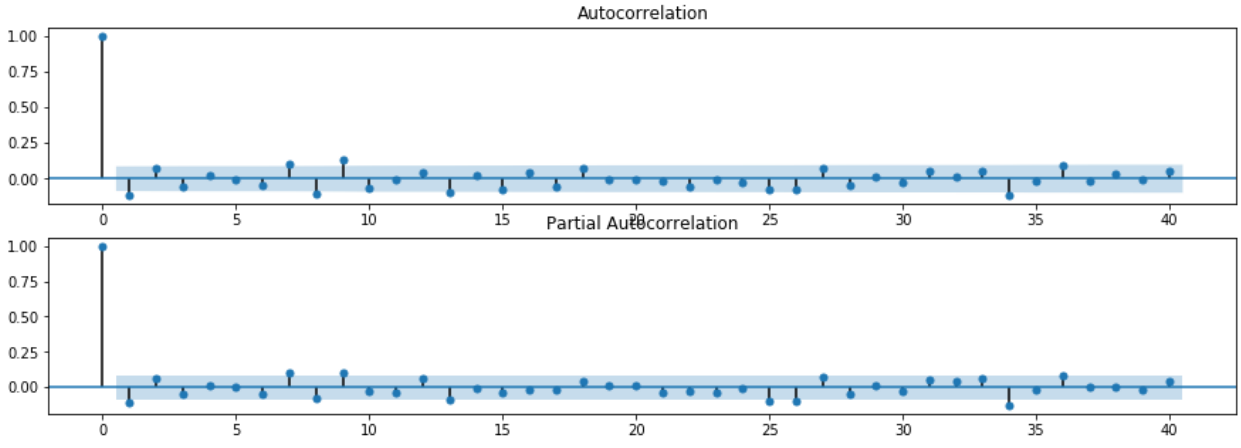
The test statistics from the Dickey-Fuller test barely crossed the 5% critical value measure. Thus, the test statistic being less than the critical value meant rejection of the null hypothesis. With the data now made stationary, it is evaluated for auto correlation using Durbin Watson Statistic method.

If the score for Durbin Watson Statistic method is between 0 and 4, where 0 depicts strong positive autocorrelation, 4 depicts strong negative autocorrelation, and 2 depicts no autocorrelation at all.

```
# Durbin Watson Statistic
sm.stats.durbin_watson(data_log_diff)
```

```
2.0882665887617886
```

As the score is 2, there is no autocorrelation in the data set. It can be observed from the plots below-



MODEL BUILDING & EVALUATION

The next step in the process is to build the time series model. The model that I chose to perform the time series analysis is, ARIMA model – Auto Regressive Integrated Moving Average model.

ARIMA forecasting revolves around a linear equation whose behavior is dependent upon the values of p, d, and q. In other words, the ARIMA model filters signal from noise and forecasts it for future point in time. An ARIMA model is defined as ARIMA (p, d, q).

An ARMA model forecasts a time series from a linear function of its past values. Time lag is defined in the AR (auto-regressive) model to forecast the future values of a time series. Moving Average is best for situations in which we have a univariate time series object. MA forecasts future values by training itself on the current and past values of a time series that are random in nature. The combined model refers to the addition of AR and MA. This model will have all the three parameters—p, d, and q—and will initially be defined as an ARIMA —ARIMA (p, d, q),

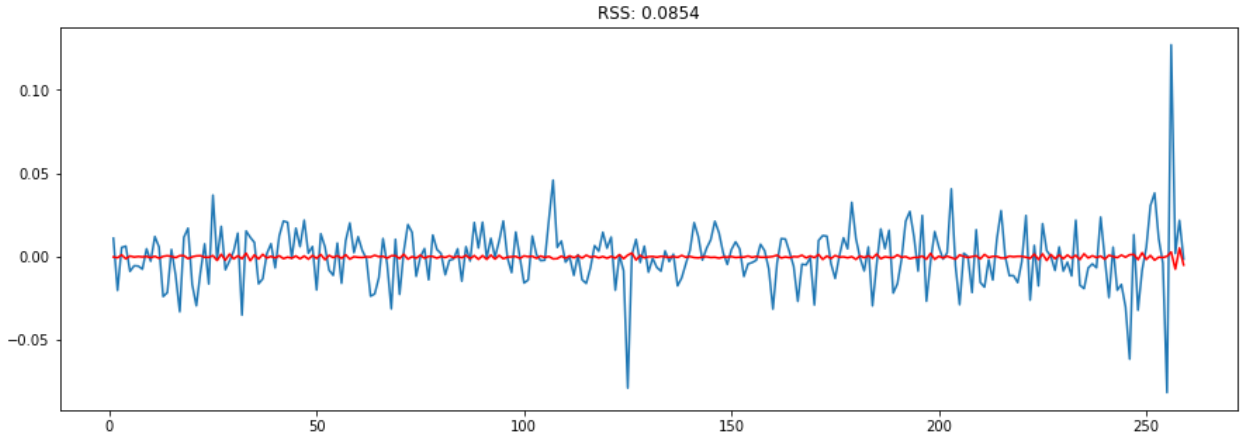
p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary.

For building this model, p, q, and d values are chosen as 1.

The model is plotted as below –



The RSS value, which is the measure of residuals is significantly less. The AIC score for the model is, -1333.321121331493. The score is extremely low, proving the goodness of the model.

The test statistics of RMSE, MFE, MAE and MAPE for the model is as below –

Root Mean Squared Error (RMSE) is the square root of residual sum of squares (RSS): **9.8666**

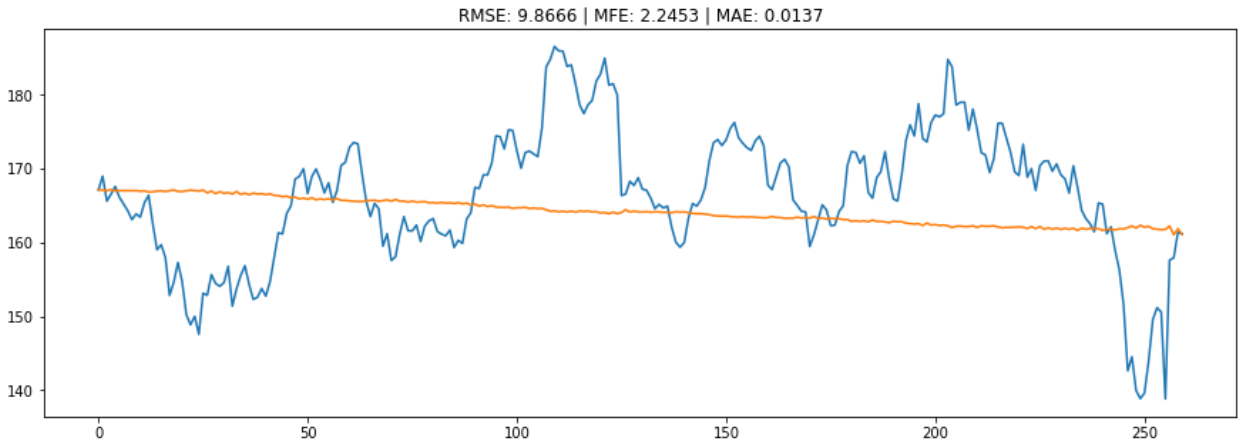
Mean Forecast Error (MFE), the mean of residuals at each time point: **2.2453**

Mean Absolute Error (MAE), the mean of absolute value of residuals at each time point: **0.0137**

Mean Absolute Percentage Error (MAPE): **0.047438918017339245**

The above values for test statistic also proves the goodness of the model.

The plot between the forecast and actual data is as below –



CONCLUSION

Around 4.7% MAPE (Mean Absolute Percentage Error) implies the model about 95.3% accurate in predicting the test set observations. From the forecast vs actual data, directionally the predictions appear to be correct.

Time series analysis of Amazon stocks thus, helps in understanding seasonality, trends, cyclicality, and randomness in closing rate data. This immensely helps in stock portfolio management for individuals and companies in making a well-informed decision.

ASSUMPTIONS

A common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. However, with the given data set, the assumption was not made, but tested to evaluate the stationary nature of the data and efforts were made to make the data stationary.

LIMITATIONS & CHALLENGES

Stock values are dependent on many socio-economic factors. The model can only indicate probability of direction of stock values. Even if the model accuracy is good, the predicted values shouldn't be considered as is without risk assumptions.

The special events like pandemic and war influence the market. Sometimes, the risk assumptions become too high, but the market becomes resilient. The challenges can be overcome by adding the special events as a feature and assess the correlation.

FUTURE USES/ADDITIONAL APPLICATIONS

The model is built using stock data for Amazon stocks; however, this can utilize to predict the closing APR for any stock. Other than stock prediction, time series analysis can be built for sales forecasting, weather forecasting, unemployment estimates, disease outbreak.

RECOMMENDATIONS

The model is built with the parameter values (1,1,1). It can be assessed to see if the model can be improved by varying these factors. Additional features about the social economic conditions, news/reviews about the company itself can be added for more accurate predictions.

IMPLEMENTATION PLAN

The machine learning models are deployed and scheduled to production in batch mode or online mode. Batch mode is a process scheduled to run daily, weekly, or monthly basis. Online mode can be made available to run on demand basis by an online application. This model can be run daily to predict for the stock trends.

ETHICAL ASSESSMENT

The stock market is affected by many known and unknown factors. In the sense of which, there should be solid risk assumptions in the prediction and the predicted value should never be deemed accurate. The agents in financial institutions should look for sustainable growth of portfolio instead of short-term profits.

REFERENCES

https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis.html

https://www.databricks.com/dataaisummit/session/challenges-time-series-forecasting

https://otexts.com/fpp2/non-seasonal-arima.html