Time Series Prediction and Forecasting of Apple.inc Stock Market Index

BY:

Vrushabh Wadnere | vw1937@rit.edu

Abstract:

Stock market analysis is a critical component of investment decision-making and economic forecasting. In this project, we conducted a comprehensive stock market analysis of Apple Inc. using time series analysis techniques to forecast future values of the closing index. Our goal was to identify the best fit model that accurately predicts the future performance of Apple Inc.'s stock.

The data used for our analysis covered a period of weeks, from March 27th, 2017 to April 28th, 2017. The data was collected and processed to obtain weekly average values of the open, high, low, close, adjusted close, and volume. We used the closing index as our response variable and applied various time series analysis techniques, including regression, exponential smoothing, and ARIMA, to develop models that can predict future values of the closing index.

Our evaluation of the models was based on overall prediction accuracy and error measures. The best fit model was identified as the one that achieved the highest accuracy and the lowest error measure. Our results showed that the best fit model achieved a baseline accuracy of 66.66% for predicting the future value of the closing index of Apple Inc.'s stock.

Our analysis highlights the potential of time series analysis in stock market analysis and provides insights that can inform investment decisions. However, it is important to note that stock market predictions are inherently uncertain and should be viewed as estimates rather than guarantees of future performance. Our findings can be used as a starting point for further analysis and to develop more sophisticated models that can provide more accurate forecasts.

Introduction:

The stock market is a vital tool for investors and businesses alike. Understanding the performance of the market and individual companies is crucial for making informed investment decisions. Moreover, it helps identify trends and patterns that can inform business decisions. Financial analysts, market researchers, and investors analyze stock market data for various reasons. Investors use the stock market index as a key indicator of the overall health and direction of the economy. They apply two main approaches to analyze stock market data: technical and fundamental analysis. Technical analysis focuses on analyzing stock charts and trends to predict future price movements. Technical analysts use various tools, such as moving averages, trend lines, and momentum indicators, to identify patterns and trends in stock prices. They believe that stock prices follow trends that repeat over time and that these patterns can be used to make predictions about future price movements. On the other hand, fundamental analysis evaluates a company's financial statements and industry trends to determine its long-term growth potential. Analysts look at factors such as revenue growth, profitability, debt levels, and competitive landscape to assess a company's current and future value. Fundamental analysis aims to identify undervalued or overvalued companies and help investors make informed investment decisions. In this report, we focus on time series analysis, which is a statistical technique used to analyze patterns and trends in time-series data. We apply various time series analysis techniques, such as regression models, exponential smoothing models, and ARIMA models, to forecast the future values of the closing index for Apple Inc.'s stock. The performance of these models is compared based on overall prediction accuracy and error measures. The report aims to demonstrate how time series analysis can be used to make informed investment decisions and identify emerging opportunities or risks in the stock market.

Dataset:

Prior to proceeding with the dataset, I had mentioned in the project proposal that I intended to use the NADAQ market index values as a predictor. However, during the implementation of the analysis, I came to the realization that incorporating the NADAQ market indexes, in addition to the Apple Inc. stock market indexes, would make the model overly complex and hinder our ability to accurately predict the forecast value for the Apple Inc. close index. Therefore, I have made the decision to move forward with the analysis without including the NADAQ index.

The dataset was obtained from Yahoo Finance! website and consists of stock market data for a company over a period of weeks, starting from March 27, 2017, and ending on April 28, 2023. The dataset contains information such as the date, opening price, highest price, lowest price, closing price, adjusted closing price, and volume of stocks traded for each week. This data can be analyzed to evaluate the company's performance in the stock market during the given timeframe. The dataset can be utilized to train and test a time series model that predicts the future values of the Apple Inc. stock market index.

index	Date	Open	High	Low	Close	Adj Close	Volume
0	2017-03-27	35.93	36.067501	35.752499	35.915001	33.705326	78646800
1	2017-04-03	35.927502	36.365002	35.762501	35.834999	33.630249	421664800
2	2017-04-10	35.900002	35.970001	35.014999	35.262501	33.092964	349942800
3	2017-04-17	35.369999	35.73	35.112499	35.567501	33.379211	356994000
4	2017-04-24	35.875	36.224998	35.794998	35.912498	33.702991	364614800
5	2017-05-01	36.275002	37.244999	36.067501	37.240002	34.948811	701406800
6	2017-05-08	37.2575	39.105	37.2575	39.025002	36.623981	693882400
7	2017-05-15	39.002499	39.162498	37.427502	38.264999	36.058956	629419600
8	2017-05-22	38.5	38.724998	38.1675	38.4025	36.188538	412906000
9	2017-05-29	38.355	38.862499	38.055	38.862499	36.622021	355011600
10	2017-06-05	38.584999	38.994999	36.505001	37.244999	35.097767	636638800
11	2017-06-12	36.435001	36.875	35.549999	35.567501	33.516983	882121600
12	2017-06-19	35.915001	36.790001	35.915001	36.57	34.461685	533012000
13	2017-06-26	36.7925	37.07	35.57	36.005001	33.92926	508240800
14	2017-07-03	36.220001	36.325001	35.602501	36.044998	33.966957	316711600
15	2017-07-10	36.0275	37.3325	35.842499	37.259998	35.111904	444353600
16	2017-07-17	37.205002	37.935001	37.142502	37.567501	35.401676	424326400
17	2017-07-24	37.645	38.497501	36.825001	37.375	35.220276	423272400
18	2017-07-31	37.474998	39.9375	37.032501	39.0975	36.843475	691234000

Fig: Weekly Dataset

A basic null value check on the dataset before using it for analysis. In the initial stage of the analysis decided not to proceed with daily apple inc. stock market data since it have many missing value for all features throughout the dataset. Due to non-recorded values for weekend and national holidays. I have implemented forward fill method to fill these missing values in the data set.

```
# Combine the new DataFrame with the original DataFrame
df = new_df.combine_first(df).reset_index()

# Fill missing values with the previous available value
df.fillna(method='ffill', inplace=True)
```

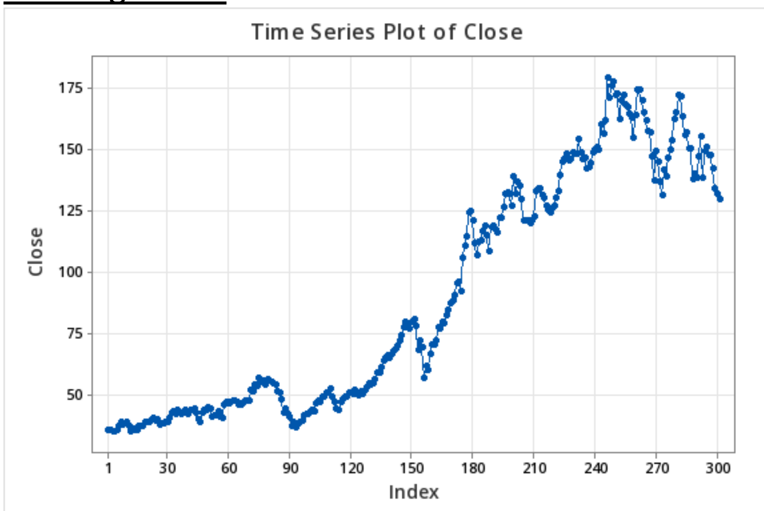
	Date	Open	High	Low	Close	Adj Close
0	2022-04-27	155.910004	159.789993	155.380005	156.570007	155.627258
1	2022-04-28	159.250000	164.520004	158.929993	163.639999	162.654694
2	2022-04-29	161.839996	166.199997	157.250000	157.649994	156.700729
3	2022-05-02	156.710007	158.229996	153.270004	157.960007	157.008881
4	2022-05-03	158.149994	160.710007	156.320007	159.479996	158.519730
5	2022-05-04	159.669998	166.479996	159.259995	166.020004	165.020355
6	2022-05-05	163.850006	164.080002	154.949997	156.770004	155.826050
7	2022-05-06	156.009995	159.440002	154.179993	157.279999	156.562668
8	2022-05-09	154.929993	155.830002	151.490005	152.059998	151.366486
9	2022-05-10	155.520004	156.740005	152.929993	154.509995	153.805298
10	2022-05-11	153.500000	155.449997	145.809998	146.500000	145.831848
11	2022-05-12	142.770004	146.199997	138.800003	142.559998	141.909805

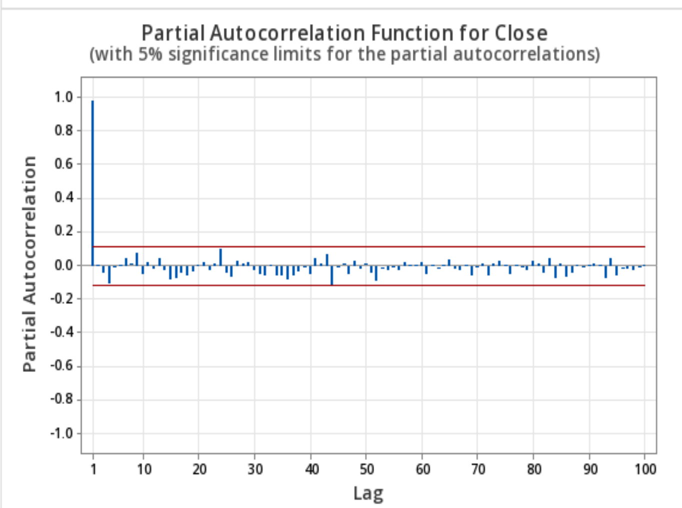
Fig: Daily dataset with missing values

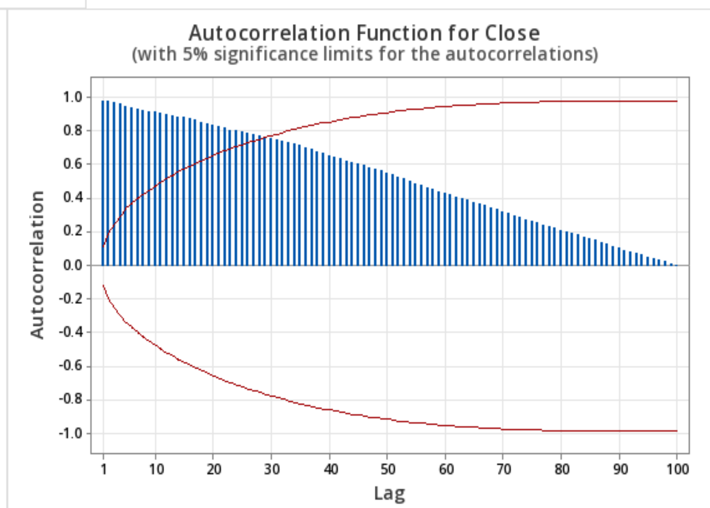
Finally, I have decided to move forward with weekly since it will have less noise which also makes it easier to identifying an underlying trend.

Analysis:

<u>Training: Close</u>







To begin with, we examine the time series plot of the closing apple stock market index over time. The plot reveals that the series displays a combination of trend to it. Specifically, it appears to experience an upward trend until roughly point 250, after which it begins to exhibit a downward trend.

It was necessary to investigate whether the series exhibited any seasonal patterns by generating ACF and PACF plots. Our analysis of the ACF plot revealed that numerous lags exceeded the 95% confidence interval, which suggests that the series is non-stationary. However, the trend identified in the ACF plot did not indicate any seasonal patterns, and this finding was corroborated by the PACF plot. Therefore, we can conclude that the series exhibits cyclic behavior rather than displaying any seasonal patterns. Notably, Lag 1 in the PACF plot surpasses the 95% confidence interval.

To determine the best model for the series, we divided the data into three separate subsets. The first subset, which contained values from April 2017 to December 2022, served as the training data. The second subset consisted of values recorded from January 2023 to February 2023 and was designated as the validation data. The final subset, containing values recorded between March 2023 and April 2023, was reserved as the testing data.

In order to create a model for this series, we have considered multiple approaches, including:

1. Regression Modeling
2. Double Exponential Smoothing
3. ARIMA Modeling

We will explore these approaches to determine which one produces the most accurate and reliable results for our analysis.

We did not consider the following methods for modeling the series due to its non-seasonal behavior:

1. Dummy Variables Methodology
2. Trigonometric Models
3. Winter Holt's Method
4. SARIMA

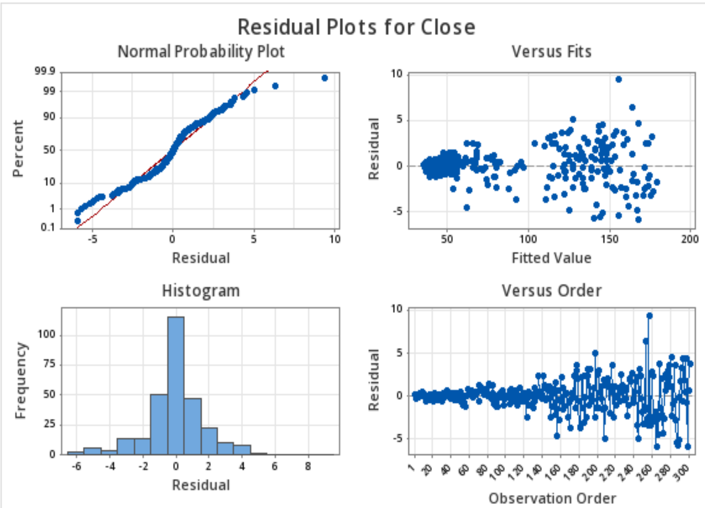
Furthermore, single exponential smoothing was also not considered as it only handles series with no trend and no seasonality, which is not applicable for our non-seasonal series with a trend.

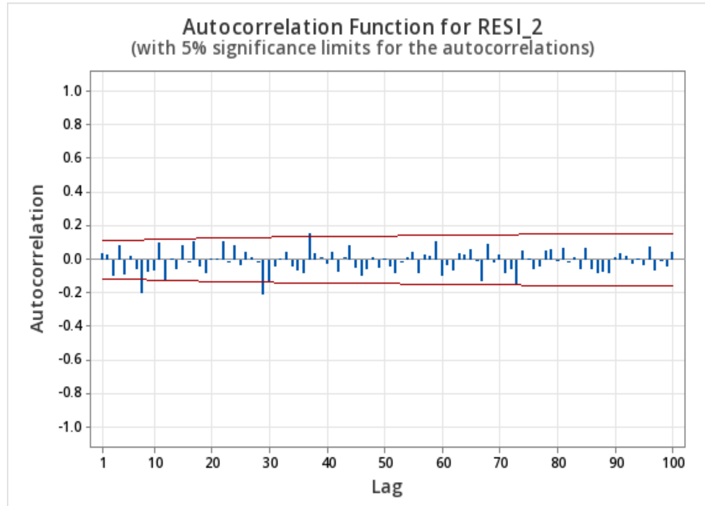
Regression Model: Close

Regression model:close

Sr. No	Terms Included	Significant Terms	MSE	S	R-sq	R Square Adjusted	R Square Predicted	Durbin Watson
1	T, O, H, L	O, H, L	4	1.94155	99.84	99.83	99.83	1.91440
2	T,T-T,O,H,L	O,H,L	4	1.94466	99.84	99.83	99.82	1.91576
3	T,T-T,T-T-T,O,H,L	O,H,L	4	1.93629	99.84	99.83	99.83	1.91283
4	T,T-T,T-T-T,T-T-T-T,O,H,L	O,H,L,T-T-T,T-T-T-T	3.7	1.92264	99.84	99.84	99.83	1.89591
5	T,T-T,T-T-T,T-T-T-T, T-T-T-T-T,O,H,L	O,H,L	3.7	1.92140	99.84	99.84	99.83	1.90116
6	T,T-T,T-T-T,T-T-T-T,O,H,Lag	O,H,L,T-T-T,T-T-T-T	3.7	1.92902	99.84	99.83	99.82	1.89247

Based on the time series plot, it was evident that the series exhibited a polynomial trend. To capture this trend, we fit a regression model that included polynomial terms of time. We added higher degree polynomial terms until they were no longer statistically significant. The initial regression model included terms for time, open, high, and low. We evaluated the models using the Durbin-Watson score and found that our best model have comparatively lower Durbin-Watson than other models. The R-squared value was also high, at 99.84%, and there was not much difference between the R-squared, R-squared adjusted, and R-squared predicted values. We also tested a fifth-degree interaction term but found it to be statistically insignificant, despite having a high Durbin-Watson score. Therefore, we decided to move forward with the fourth-degree model, which had overall better performance than other models. Additionally, we added a lag term for high values at PACF lag 1, but found it to be insignificant and discarded it.





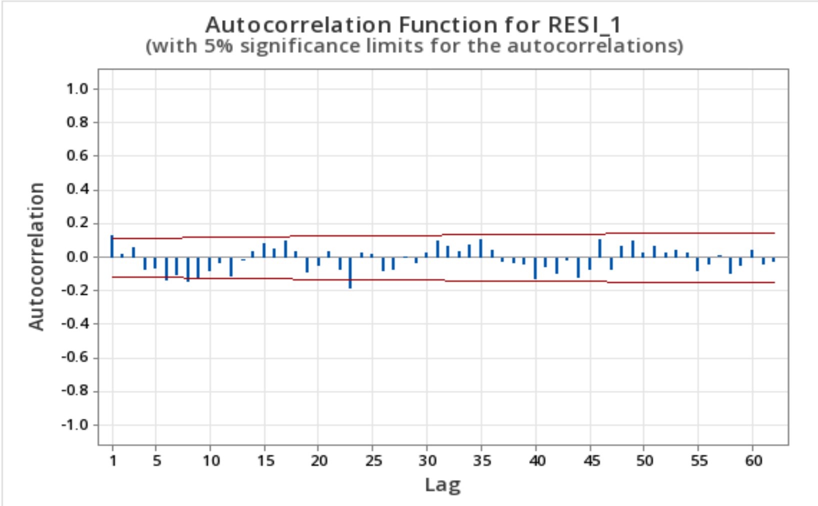
Our goal was to enhance the accuracy of our model, and to do so, we needed to obtain a model with a low mean square error and high prediction accuracy. To achieve this, we introduced higher degree polynomial terms into the model, which helped to decrease the mean square error and correlation between the residuals. We discovered that adding polynomial terms up to the fourth degree for time had a statistically significant impact. In terms of the autocorrelation function of the residuals for this regression model, we observed that there was no notable correlation within the residuals, with most lags falling within the 95% confidence intervals, with only a few exceptions.

<u>Double Exponential Model: Close</u>

Double Exponential Smoothing: close

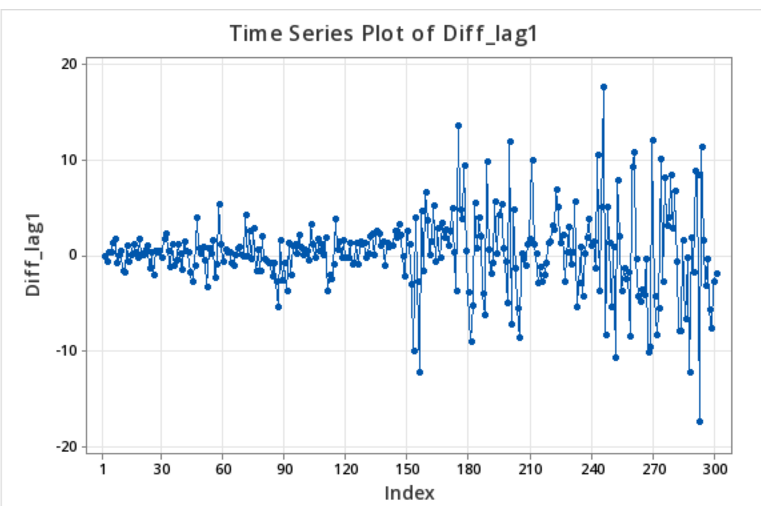
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	6.7498	5.5198	56.9566
2	0.2	0.1	6.6707	5.3294	52.0688
3	0.2	0.50	6.5607	5.1985	49.2385
4	0.3	0.1	5.4557	4.5125	38.6972
5	0.5	0.1	4.1899	3.5826	26.1169
6	0.8	0.1	3.4225	2.9838	20.3025

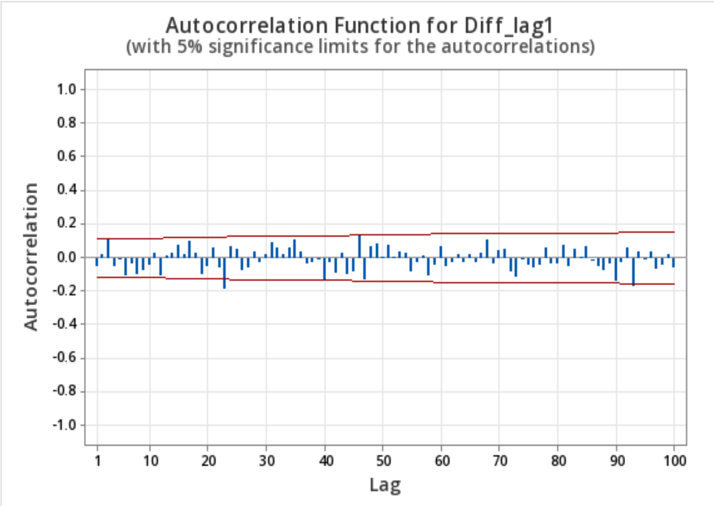
Our second modeling approach involves utilizing Double Exponential Smoothing, wherein we use the Level and Trend levers to construct the most effective model. Initially, we set the level and trend to 0.2, which resulted in an unsatisfactory MSD score of 56.9566, demonstrating that our model was inferior to the baseline regression model. To enhance the model, we attempted to reduce the Trend while maintaining the Level constant, and we observed a gradual decline in MSD. We also tried using smaller trend values, which led to additional decreases in MSD. Using the table, we identified model 6 as our best double exponential smoothing model, with a Level of 0.8 and a Trend of 0.1. We concluded that altering the level or trend did not have a significant impact on the MAD, MAPE, and MAD metrics, so we decided to stop the process.

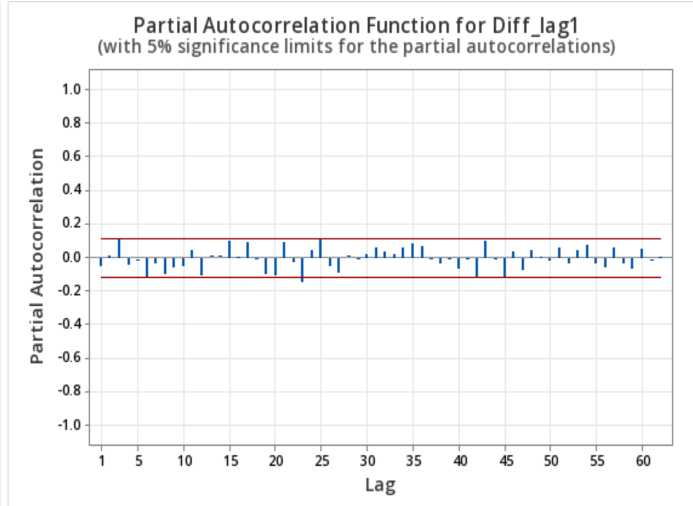


We then analyzed the autocorrelation function (ACF) of model 6 from the Double Exponential Smoothing approach. The ACF plot indicated that, except for a few lags, all other lags were within the 95% confidence interval, indicating that the residuals of the double exponential smoothing model did not possess excessive autocorrelation. However, as our objective is to attain high precision and the mean squared error (MSE) of 20.3025 is significantly poorer compared to our regression model, we decided to abandon this approach.

ARIMA Model: Close







We proceeded with the ARIMA modeling technique by computing the first-order difference of the time series, which served to remove the trend. The resulting plot of the first-order difference (Diff1) indicates that the Diff1 series could potentially be a white noise process. To confirm this hypothesis, we analyzed the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the Diff1 series, and the results provided evidence that the series is indeed a white noise process.

After analyzing the ACF and PACF plots, we major find any lags outside the 95% confidence interval. To find the best fit, we tested different combinations of AR and MA terms for the ARIMA model. We also examined additional MA values while keeping the AR term constant to evaluate whether any of the models could accurately represent the series.

Sr.	**AR**	**Difference**	**MA**	**Significant Terms**	**MSE**	**Goodness of Fit**	**SS**
1	1	1	1	-	17.4642	Fail	5186.88
2	1	1	2	AR1,MA1	17.1758	Fail	5084.04
3	1	1	4	AR1,MA1,MA4	17.1479	Fail	5041.48
3	1	1	5	AR1,MA1	17.2049	Fail	5041.03
4	4	1	5	AR1,AR2,AR3,AR4,MA1,MA2,MA3,MA4	16.6138	Fail	4818.01
5	3	1	3	MA3	17.0532	Fail	4996.59

Similar to our Double Exponential Smoothing approach, the ARIMA models also had notably higher mean square error (MSE) values compared to the regression model. Moreover, all the ARIMA models failed the goodness of fit test. Therefore, we concluded that the ARIMA modeling technique was not suitable for our data and discarded it.

<u>Validation: close</u>

Based on our analysis, we determined that the best model for predicting the Apple Inc. stock market index was the regression model. Both the Double Exponential Smoothing and ARIMA techniques were discarded due to their high mean squared error. Below are the predicted values using our final regression model.

Regression model

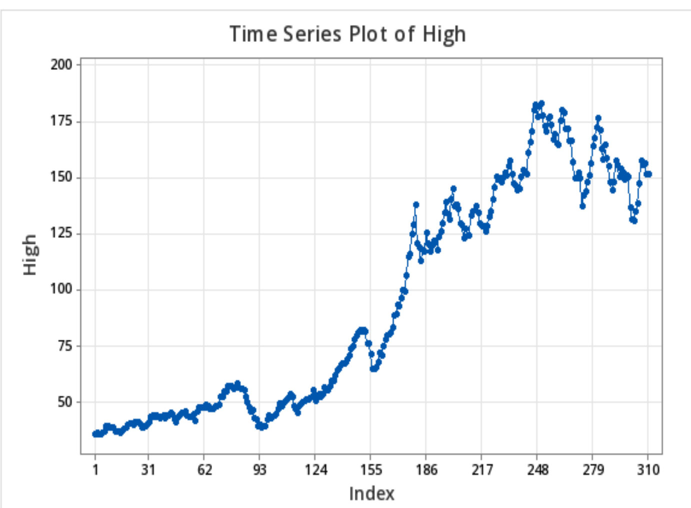
Validation (Jan-Feb): close									
Time	Year	Month	Actual Values	FOR_REG	LOW_REG	UPP_REG	Actual in PI_REG	MAE_REG	MSE_REG
302	2023	Jan	129.620	125.011	121.043	128.978	No	4.609	21.242881
303	2023	Jan	134.760	131.262	127.287	135.236	Yes	3.498	12.236004
304	2023	Jan	137.870	135.787	131.770	139.804	Yes	2.083	4.338889
305	2023	Jan	145.930	143.921	139.868	147.974	Yes	2.009	4.036081
306	2023	Jan	154.500	150.379	146.261	154.497	No	4.121	16.982641
307	2023	Feb	151.010	149.645	145.4941	153.797	Yes	1.365	1.863225
308	2023	Feb	152.550	153.081	148.873	157.288	Yes	0.531	0.281961
309	2023	Feb	146.710	144.959	140.756	149.162	Yes	1.751	3.066001
310	2023	Feb	151.030	144.906	140.694	149.119	No	6.124	37.503376

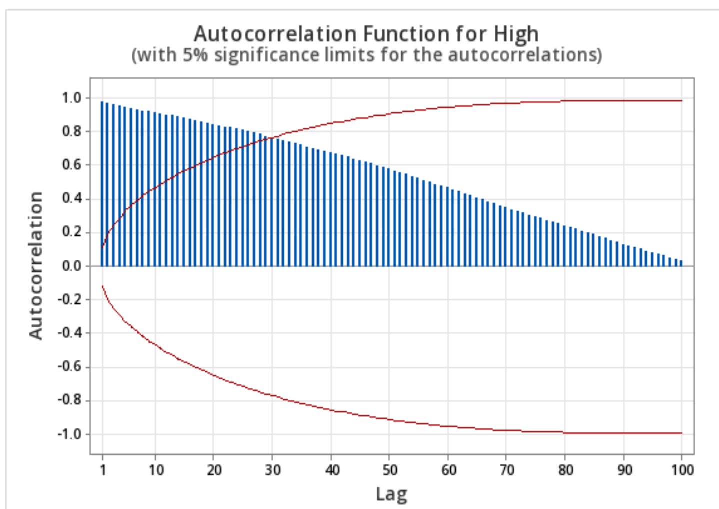
The regression model yielded a prediction accuracy of 66.66%, and the Mean Absolute Error was calculated to be 3.7. Overall, the regression model showed satisfactory precision. Considering the high MSE obtained by both ARIMA and Double Exponential Smoothing techniques, we have decided to proceed with the regression model.

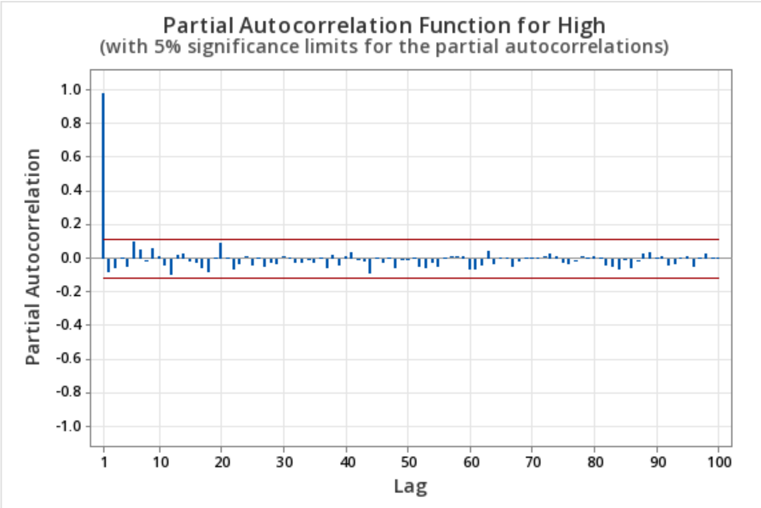
Now that we have trained and validated our best model which includes the terms open, high, and close, the next step is to test it. In order to do this, we need to find optimal models for each of these terms. To accomplish this, we will train all of these models using data from April 2017 to February 2023 and validate them using data from March 2023 to April 2023. Since the open term may be dependent on the high and low terms, we will begin by analyzing the high series.

Training: High

We begin by examining the time series plot for High. Similar to Close, the plot suggests that the series exhibits a combination of trends. Specifically, the series appears to follow an upward trend until time 248, after which it begins to decrease.







Upon analyzing the autocorrelation function (ACF) of the High series, we observe that many of the lags lie above the 95% confidence interval, indicating that the series is not stationary. The trend in the ACF plot doesn't seem to be seasonal, as confirmed by the partial autocorrelation function (PACF) plot. Therefore, we can conclude that the behavior of the series is not seasonal and is cyclic in nature. Additionally, the PACF shows that Lag 1 passes the 95% confidence interval.

(High) Regression model

Sr. No	Terms Included	Significant Terms	MSE	S	R-sq	R Square Adjusted	R Square Predicted	Durbin Watson
1	T,T-T,T-T-T,T-T-T-T,T-T-T-T-T,Lag1,Lag2	T,T-T,T-T-T,T-T-T-T,T-T-T-T-T,Lag1,Lag2	12	3.47691	99.51	99.50	99.48	2.03319

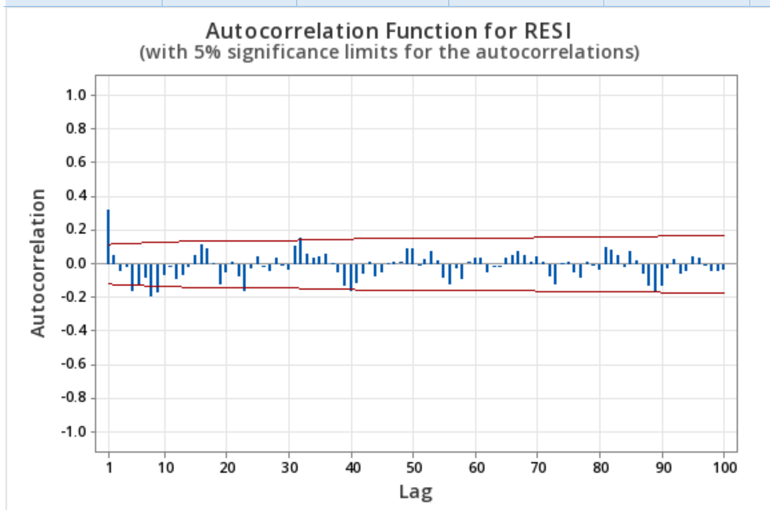
Based on the time series plot, we observed a polynomial trend in the high series similar to the close series. Thus, we utilized the same three models used for the close series. We fitted a regression model that included polynomial terms of Time, gradually adding higher degree polynomial terms until they were no longer significant. Additionally, as we observed high correlation at Lag 1 in the PACF plot, we included Lag 1 of high to the model to reduce autocorrelation between residuals. However, despite adding Lag 1, there was still significant autocorrelation between residuals, prompting the inclusion of Lag 2 as well. The Durbin-Watson score of this model was good at 2.03319, and the R-square value was impactful at 99.51%.

Double Exponential model: High

Our next modeling approach is Double Exponential Smoothing, and we'll be using the same Level and Trend levers that we used before to fit the “Close “model.

High Double Exponential Smoothing

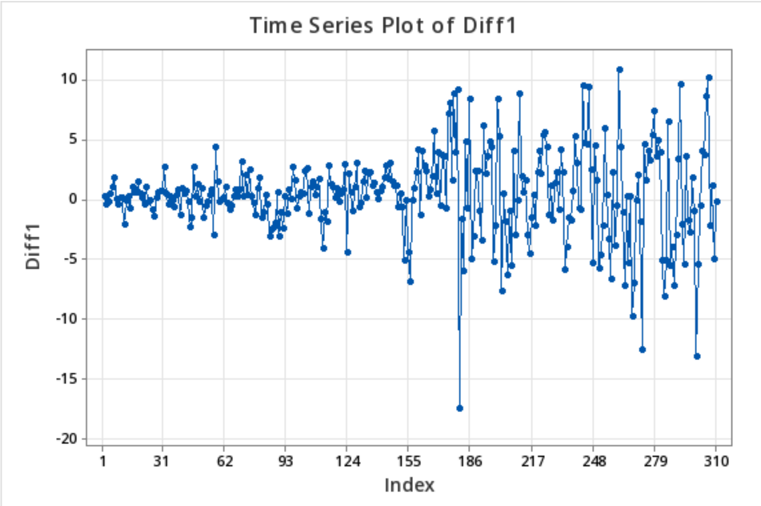
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	6.3925	5.5890	58.5680
2	0.4	0.4	4.2931	3.9357	31.3519
3	0.8	0.1	3.0729	2.8210	17.0757
4	0.8	0.2	3.0937	2.8384	17.4221

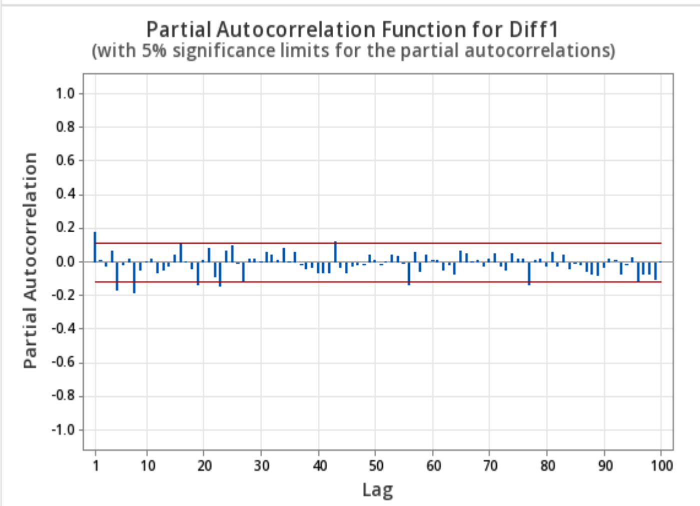


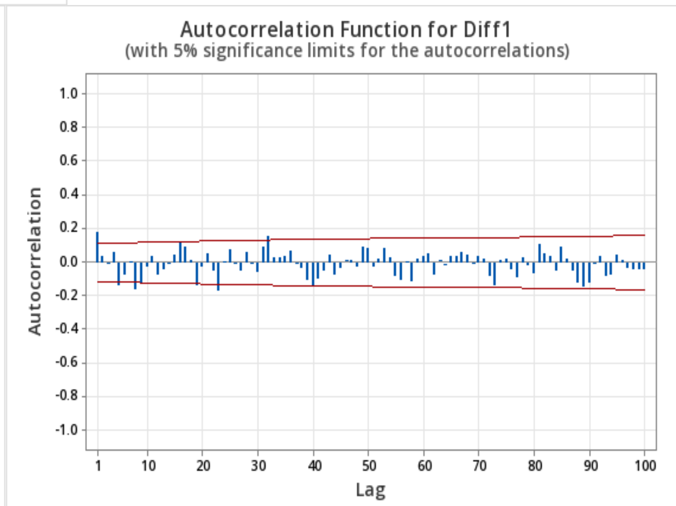
Next, we examine the autocorrelation function (ACF) plot of model 3 obtained from the Double Exponential Smoothing approach. The plot shows that most of the lags are within the 95% confidence interval, indicating that the residuals of the model do not exhibit significant autocorrelation.

ARIMA Model:High

Moving to the ARIMA modeling technique, we first calculate the first difference (Diff1) of Lag 1 from the time series to eliminate the trend. The plot of the Diff1 series shows that it might be white noise. To confirm this, we examined the ACF and PACF plots of Diff1, which revealed a few lags above the 95% confidence interval.







Now observing the above ACF and PACF plots we did have lag 1 and lag 5 outside the 95% confidence interval for both the plots I. So, the possible combinations of ARIMA are (1,1,1), (1,1,5), (5,1,1) and (5,1,5). The final ARIMA model is given below.

Sr.	AR	Difference	MA	Significant Terms	MSE	Goodness of Fit	SS
1	1	1	1	-	12.8527	Fail	3932.94
2	5	1	1	AR5	12.5813	Fail	3799.57

Since it fails to satisfy goodness of fit test we will be discarding this models.

Validate: High(March and April)

As the MSE values for both regression and double exponential model is high, we proceeded to calculate the forecasted values for each approach from March to April 2023 for validation purposes.

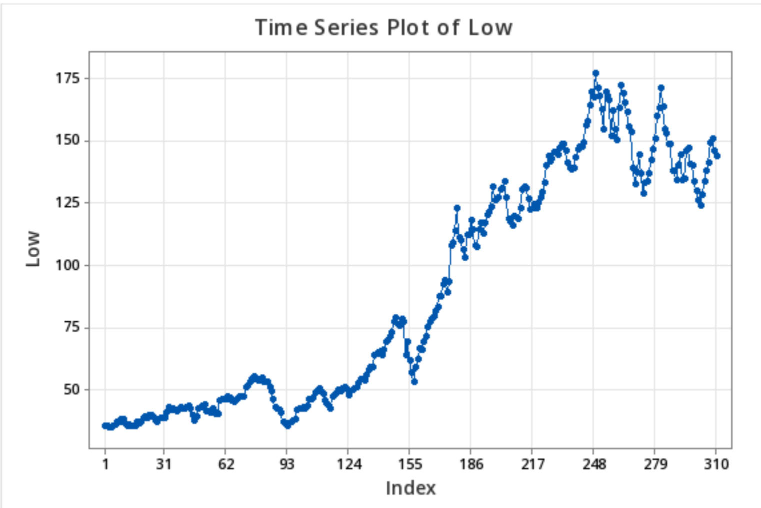
Validation (Mar-Apr): Regression: High									
Time	Year	Month	Actual Values	FOR_REG	LOW_REG	UPP_REG	Actual in PI_REG	MAE_REG	MSE_REG
311	2023	Mar	156.300	150.039	142.756	157.323	Yes	6.261	39.200121
312	2023	Mar	156.740	148.822	141.487	156.157	No	7.918	62.694724
313	2023	Mar	162.140	147.647	140.258	155.036	No	14.493	210.047049
314	2023	Mar	165.000	146.553	139.106	154.00	No	18.447	340.291809
315	2023	Apr	166.840	145.542	138.031	153.053	No	21.298	453.604804
316	2023	Apr	166.320	144.606	137.024	152.188	No	21.714	471.497796
317	2023	Apr	168.160	143.736	136.077	151.396	No	24.424	596.531776
318	2023	Apr	169.850	142.925	135.180	150.669	No	26.925	724.955625
319	2023	Apr	169.850	141.964	134.121	149.806	No	27.886	777.628996

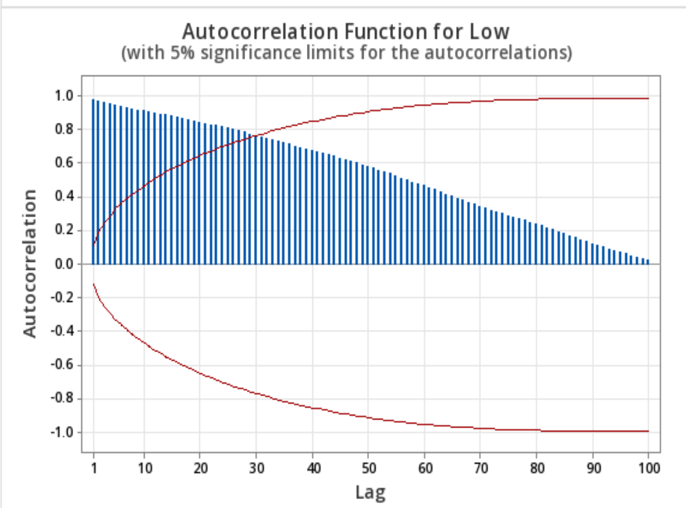
Validation (Mar-Apr): Double Exponential Model: high									
Time	Year	Month	Actual Values	FOR_DE	LOW_DE	UPP_DE	Actual in PI_DE	MAE_DE	MSE_DE
311	2023	Mar	156.300	151.707	144.796	158.618	yes	4.593	21.095649
312	2023	Mar	156.740	151.954	142.682	161.227	yes	4.786	22.905796
313	2023	Mar	162.140	152.202	140.392	164.011	yes	9.938	98.763844
314	2023	Mar	165.000	152.449	138.018	166.88	yes	12.551	157.527601
315	2023	Apr	166.840	152.696	135.599	169.794	yes	14.144	200.052736
316	2023	Apr	166.320	152.943	133.153	172.734	yes	13.377	178.944129
317	2023	Apr	168.160	153.191	130.689	175.692	yes	14.969	224.070961
318	2023	Apr	169.850	153.438	128.214	178.662	yes	16.412	269.353744
319	2023	Apr	169.850	153.685	125.731	181.639	yes	16.165	261.307225

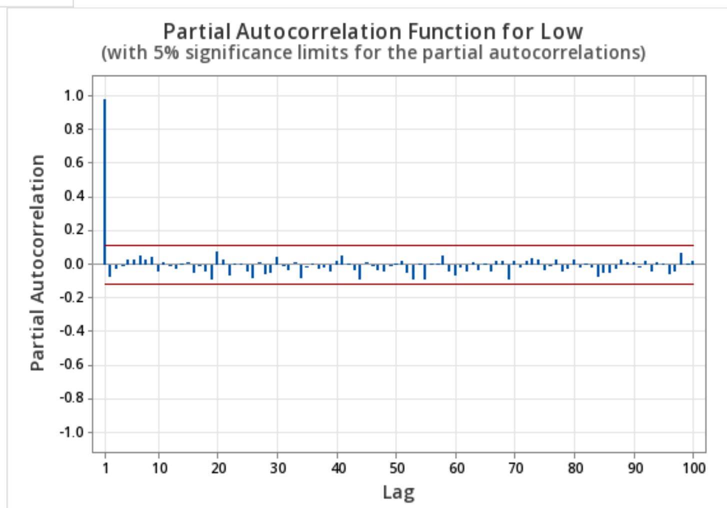
The regression model only had an 11% prediction accuracy, while the double exponential model had a 100% prediction accuracy. However, the double exponential model has wider prediction bounds, which goes against our goal of achieving high precision. On the other hand, the regression model failed to accurately forecast future values. After careful consideration, we have decided to proceed with the double exponential model as our optimal model for High.

Training : Low

We can observe that the time series plot for low is similar to that of close, indicating a combination of trends. The series shows an increasing trend until time 248 and then starts to decrease.







The ACF plot we can see that several lags are above the 95% confidence interval, indicating non-stationarity of the series. The PACF plot indicates that the series does not have a seasonal trend, but it does have a cyclic nature as the behavior is not uniform across time. The PACF also shows that Lag 1 is statistically significant at the 95% confidence level.

Regression Model: Low

(low) Regression model								
Sr. No	Terms Included	Significant Terms	MSE	S	R-sq	R Square Adjusted	R Square Predicted	Durbin Watson
1	T,T-T,T-T-T,T-T-T-T,T-T-T-T-T,Lag1,Lag2	T,T-T,T-T-T,T-T-T-T,T-T-T-T-T,Lag1,Lag2	15	3.87318	99.31	99.30	99.27	2.00782

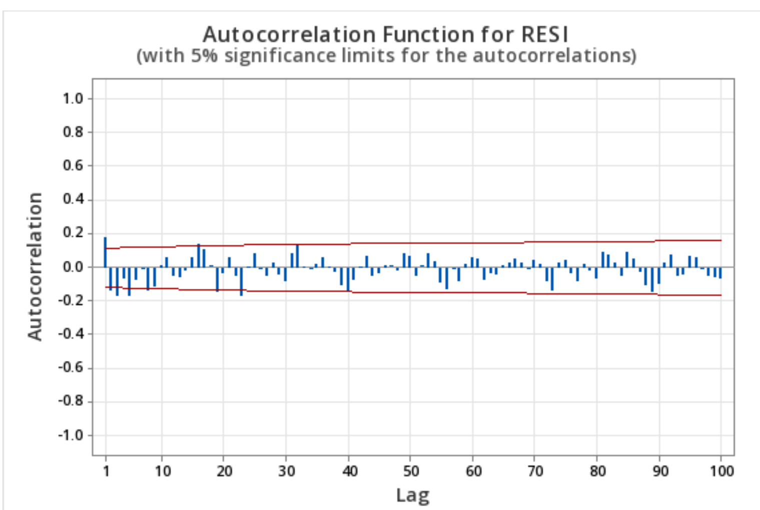
We applied the same three modeling approaches as we did for the 'Close' series since the 'Low' series also had a similar trend. We first fitted a regression model that included polynomial terms of Time, as the time series plot suggested that the series had some polynomial trend to it. We continued adding higher degree polynomial terms of Time in the model until they were no longer significant. Since we observed a high correlation at Lag 1 in the PACF plot, we added Lag 1 of 'Open' to the model to reduce the autocorrelation between residuals. However, even after adding Lag 1, we observed significantly high autocorrelation between the residuals. Therefore, we decided to add Lag 2 as well. The Durbin-Watson score of this model was good at 2.00782, and the R-square value was impactful at 99.31%. The table above shows the final regression model that we obtained.

Double Exponential Smoothing: Low

Our next modeling approach is Double Exponential Smoothing, and we'll be using the same Level and Trend levers that we used before to fit the “Close “model.

Low Double Exponential Smoothing

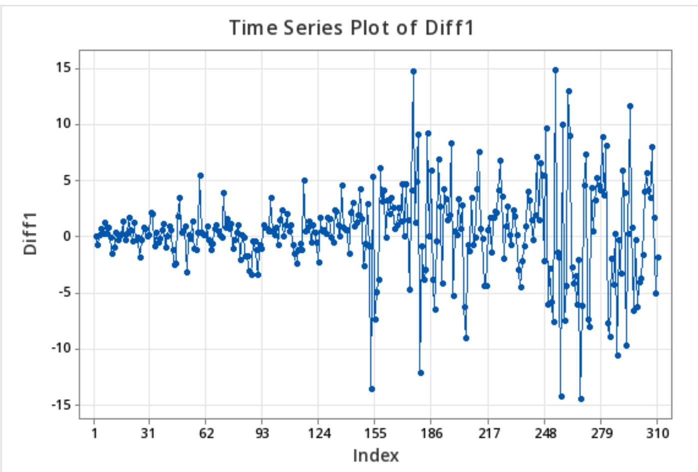
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	6.3925	5.5890	58.5680
2	0.4	0.4	4.2931	3.9357	31.3519
3	0.8	0.4	3.1179	2.8367	17.7909

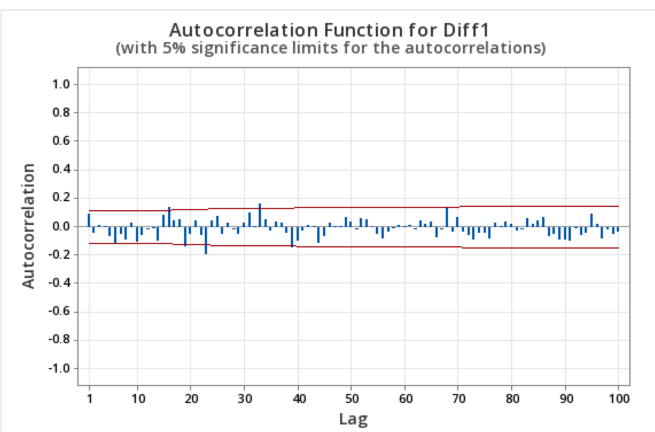


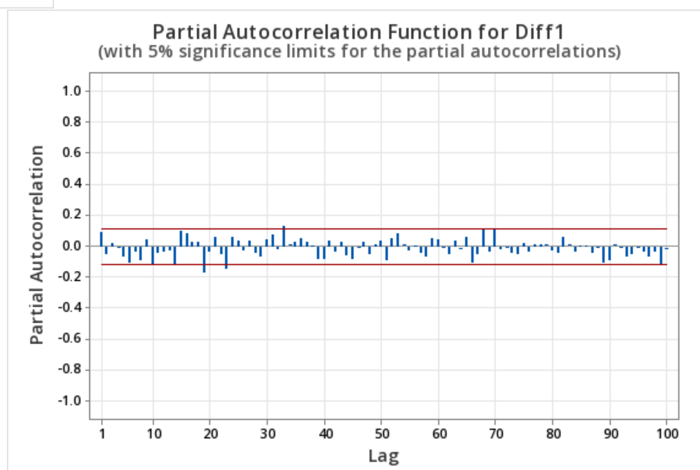
The ACF plot of the residuals from the third double exponential smoothing model shows that few of the lags are outside the 95% confidence interval, indicating that the model residuals do not exhibit significant autocorrelation except for a few lags.

<u>ARIMA Model : Low</u>

Moving to the ARIMA modeling technique, we first calculate the first difference (Diff1) of Lag 1 from the time series to eliminate the trend. The plot of the Diff1 series shows that it might be white noise. To confirm this, we examined the ACF and PACF plots of Diff1, which revealed a few lags above the 95% confidence interval.







After analyzing the ACF and PACF plots, we major find some lags outside the 95% confidence interval. To find the best fit, we tested different combinations of AR and MA terms for the ARIMA model. We also examined additional MA values while keeping the AR term constant to evaluate whether any of the models could accurately represent the series.

Sr.	**AR**	**Difference**	**MA**	**Significant Terms**	**MSE**	**Goodness of Fit**	**SS**
1	0	1	1	MA1	15.9844	FAIL	4907.22
2	1	1	1	MA1	16.0075	FAIL	4898.31
3	1	1	3	AR1,MA1	15.9953	FAIL	4862.59
4	1	1	5	AR1,MA1	15.9480	FAIL	4816.31
5	4	1	5	AR1,AR3,AR4,MA1,MA3,MA4	15.4559	FAIL	4621.31

We decided to discard this model since it fails to satisfy the goodness of fit test.

Validate: Low

As the MSE values for both regression and double exponential model is high, we proceeded to calculate the forecasted values for each approach from March to April 2023 for validation purposes

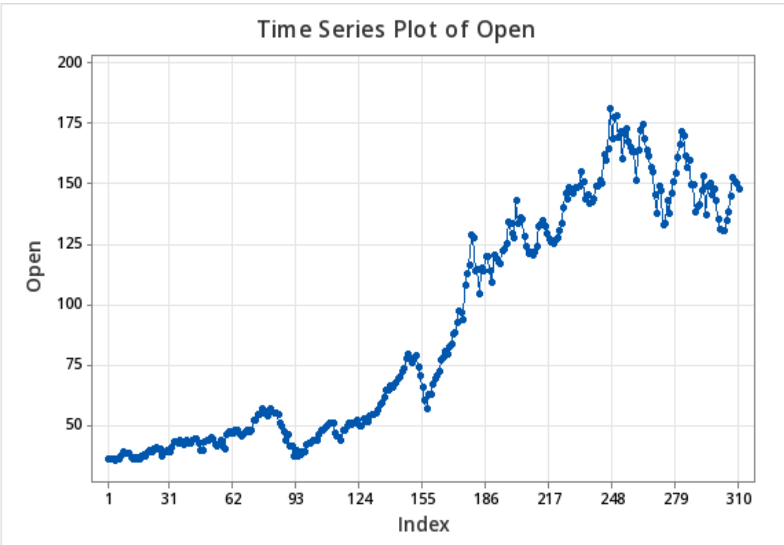
Validation (Mar-Apr): Regression: low									
Time	Year	Month	Actual Values	FOR_REG	LOW_REG	UPP_REG	Actual in PI_REG	MAE_REG	MSE_REG
311	2023	Mar	156.300	142.240	134.101	150.379	NO	14.06	197.6836
312	2023	Mar	156.740	140.765	132.577	148.953	NO	15.975	255.200625
313	2023	Mar	162.140	139.456	131.212	147.700	NO	22.684	514.563856
314	2023	Mar	165.000	138.289	129.982	146.596	NO	26.711	713.477521
315	2023	Apr	166.840	137.242	128.865	145.620	NO	29.598	876.041604
316	2023	Apr	166.320	136.297	127.841	144.753	NO	30.023	901.380529
317	2023	Apr	168.160	135.440	126.897	143.982	NO	32.72	1070.5984
318	2023	Apr	169.850	134.658	126.021	143.295	NO	35.192	1238.47686
319	2023	Apr	169.850	133.941	125.200	142.681	N0	35.909	1289.45628

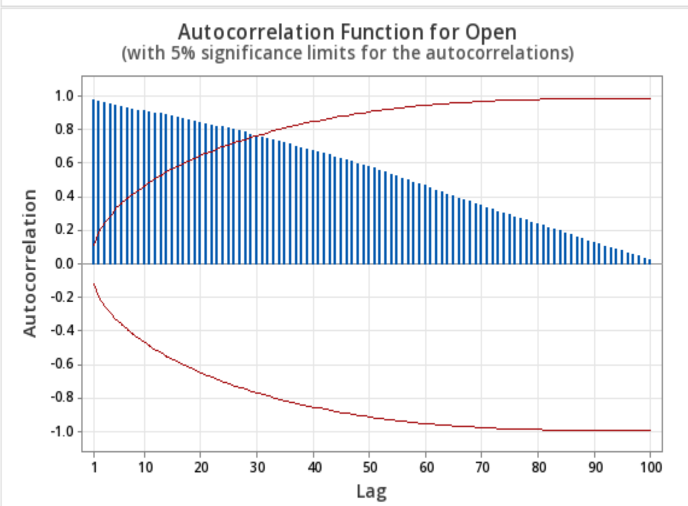
Validation (Mar-Apr): Double Exponential Model: low									
Time	Year	Month	Actual Values	FOR_DE	LOW_DE	UPP_DE	Actual in PI_DE	MAE_DE	MSE_DE
311	2023	Mar	156.300	144.058	137.553	150.563	NO	12.242	149.866564
312	2023	Mar	156.740	144.42	132.557	156.284	NO	12.32	151.7824
313	2023	Mar	162.140	144.783	127.425	162.141	YES	17.357	301.265449
314	2023	Mar	165.000	145.145	122.255	168.035	YES	19.855	394.221025
315	2023	Apr	166.840	145.507	117.069	173.945	YES	21.333	455.096889
316	2023	Apr	166.320	145.869	111.875	179.863	YES	20.451	418.243401
317	2023	Apr	168.160	146.231	106.677	185.786	YES	21.929	480.881041
318	2023	Apr	169.850	146.594	101.475	191.712	YES	23.256	540.841536
319	2023	Apr	169.850	146.956	96.272	197.64	YES	22.894	524.135236

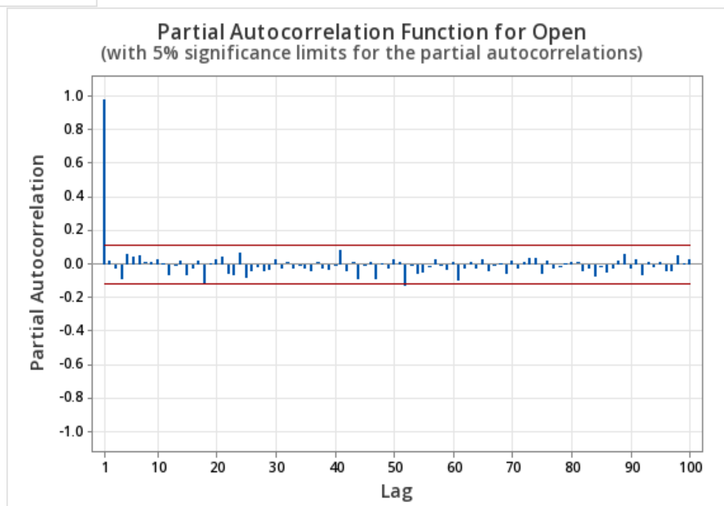
While the double exponential model had a 88% prediction accuracy. However, the double exponential model has wider prediction bounds, which goes against our goal of achieving high precision. On the other hand, the regression model drastically failed to forecast future values. After careful consideration, we have decided to proceed with the double exponential model as our optimal model for low.

Training: Open:

The next series we need to model is the Open series. Its time series plot exhibits a similar pattern to the Close series and suggests that the series has a combination of trends. The plot also shows that the series follows an upward trend until time 248 and then starts to decrease.







Upon analyzing the ACF plot, we can observe that several lags are beyond the 95% confidence interval, indicating that the series is not stationary. The ACF plot does not indicate any seasonality trend, which is verified by the PACF plot, leading us to conclude that the series behavior is cyclic rather than seasonal. Furthermore, the PACF plot shows that Lag 1 exceeds the 95% confidence interval.

Regression Model:Open

Open - Regression model - Training (Sep-Nov)

Sr.	Terms Included	Significant Terms	MSE	S	R-sq	R Square Adjusted	R Square Predicted	Durbin Watson
1	**T, H, L**	**H, L**	**4**	**2.09776**	**99.81**	**99.81**	**99.80**	**2.25624**

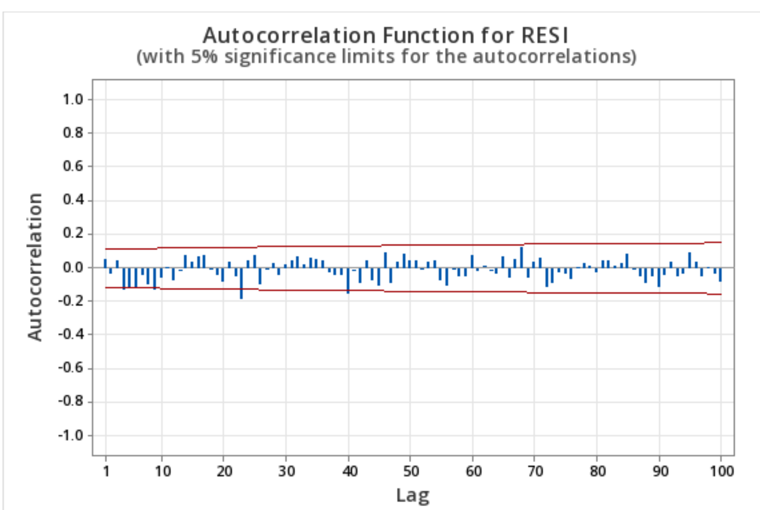
We used the same three models as before to model the series for Open, as it had a similar pattern to the series for Close. The time series plot suggested that the series had a polynomial trend, so we fitted a regression model that included polynomial terms of Time. We added higher degree polynomial terms of Time in the model until they were no longer significant. Since Open likely had a dependence on High and Low, we included those terms in the model as well. Despite the fact that the time series seemed to follow more than a linear trend, none of the Time interaction terms were found to be significant, so we reduced the model to just the Time term. The Durbin-Watson score of this model was high at 2.25624, indicating some autocorrelation between residuals. The R-square value was impactful at 99.81%. The final regression model in table above.

Double Exponential Model:Open

Our next modeling approach is Double Exponential Smoothing, and we'll be using the same Level and Trend levers that we used before to fit the "Close "model.

Open-Double Exponential Smoothing

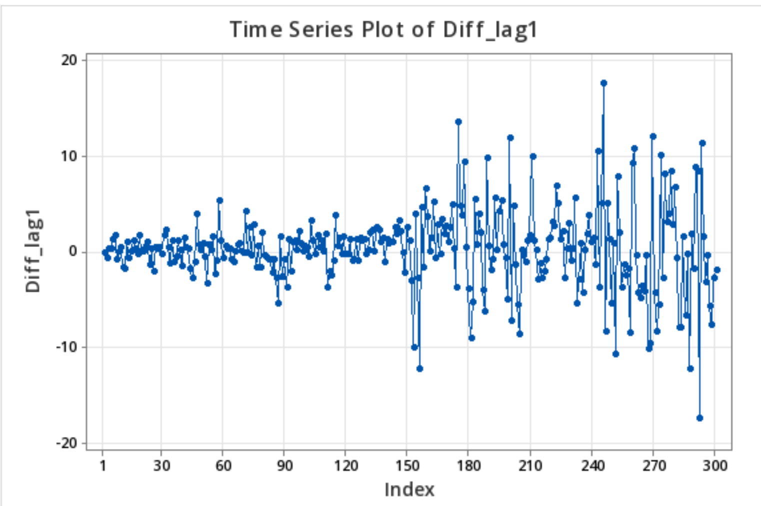
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	6.6753	5.5814	58.8907
2	0.4	0.4	4.7267	4.2014	34.7135
3	0.8	0.4	3.5800	3.2547	24.5718
4	0.8	0.2	3.5725	3.2362	23.1081

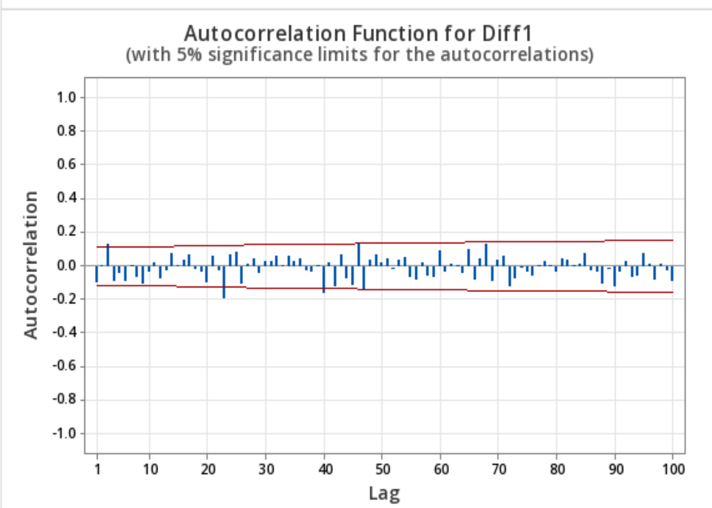


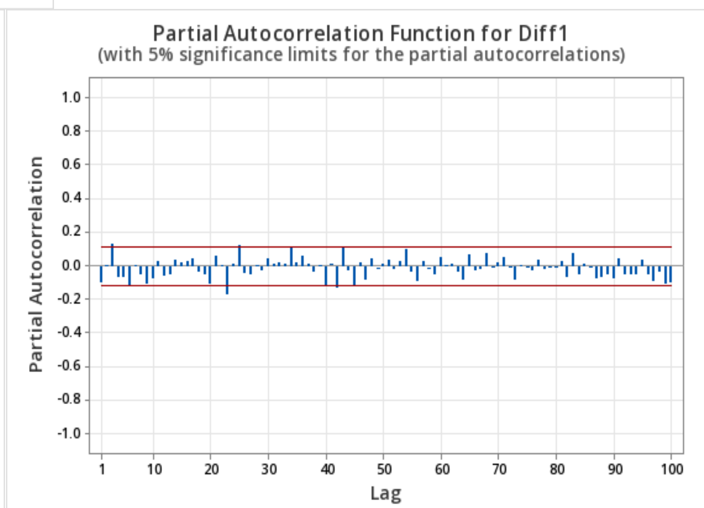
We can now take a look at the ACF plot of model 4, which is obtained using the Double Exponential Smoothing approach. The plot shows that most of the lags are within the 95% confidence interval, indicating that the residuals of the model do not exhibit significant autocorrelation.

ARIMA Model: Open:

Moving to the ARIMA modeling technique, we first calculate the first difference (Diff1) of Lag 1 from the time series to eliminate the trend. The plot of the Diff1 series shows that it might be white noise. To confirm this, we examined the ACF and PACF plots of Diff1, which revealed a few lags above the 95% confidence interval.



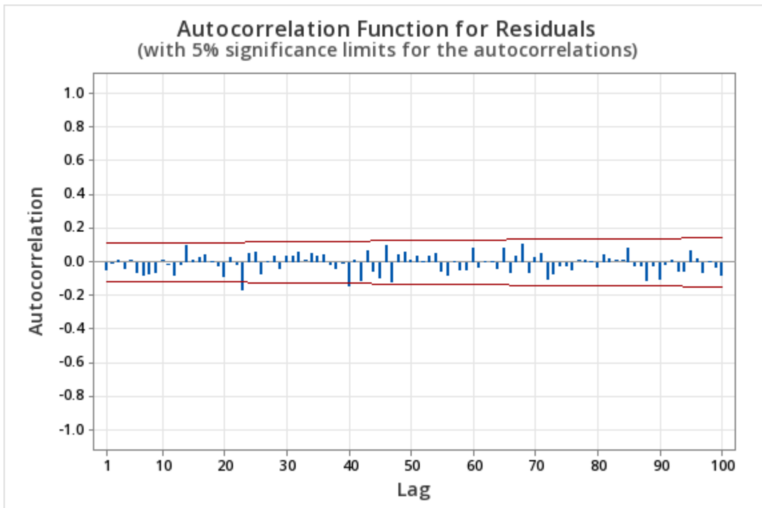




Based on the ACF and PACF plots, we can see that lag 3 are outside the 95% confidence interval for both plots. Therefore, we can consider the possible combinations of ARIMA models to be (1,1,1), (3,1,3). Below table represent the best model.

Open ARIMA

Sr.	AR	Difference	MA	Significant Terms	MSE	Goodness of Fit	SS
1	3	1	3	AR1,AR2,AR3,MA1,MA2,MA3	18.9050	PASS	5709.32



We proceed to analyze the ACF plot of the chosen ARIMA model and note that most of the lags are within the 95% confidence interval except, implying that the residuals of the ARIMA model do not display considerable autocorrelation.

Validate: OPEN:

So, to validate the models, we computed the forecasted values for March-April 2023 for all three modeling approaches.

Validation (Mar-Apr): Regression: OPEN									
Time	Year	Month	Actual Values	FOR_REG	LOW_REG	UPP_REG	Actual in PI_REG	MAE_REG	MSE_REG
311	2023	Mar	156.300	151.344	147.182	155.507	NO	4.956	24.561936
312	2023	Mar	156.740	151.627	147.465	155.789	NO	5.113	26.142769
313	2023	Mar	162.140	157.456	153.293	161.619	NO	4.684	21.939856
314	2023	Mar	165.000	159.845	155.686	164.004	NO	5.155	26.574025
315	2023	Apr	166.840	163.425	159.244	167,607	YES	3.415	11.662225
316	2023	Apr	166.320	162.247	158.077	166.418	YES	4.073	16.589329
317	2023	Apr	168.160	165.137	160.945	169.329	YES	3.023	9.138529
318	2023	Apr	169.850	165.530	161.362	169.699	NO	4.32	18.6624
319	2023	Apr	169.850	167.766	163.545	171.987	YES	2.084	4.343056

Validation (Mar-Apr): Double Exponential Model: OPEN									
Time	Year	Month	Actual Values	FOR_DE	LOW_DE	UPP_DE	Actual in PI_DE	MAE_DE	MSE_DE
311	2023	Mar	156.300	149.5	141.571	157.428	YES	6.8	46.24
312	2023	Mar	156.740	150.335	139.697	160.972	YES	6.405	41.024025
313	2023	Mar	162.140	151.17	137.621	164.718	YES	10.97	120.3409
314	2023	Mar	165.000	152.005	135.449	168.561	YES	12.995	168.870025
315	2023	Apr	166.840	152.84	133.226	172.455	YES	14	196
316	2023	Apr	166.320	153.675	130.971	176.38	YES	12.645	159.896025
317	2023	Apr	168.160	154.51	128.697	180.324	YES	13.65	186.3225
318	2023	Apr	169.850	155.346	126.409	184.283	YES	14.504	210.366016
319	2023	Apr	169.850	156.181	124.111	188.25	YES	13.669	186.841561

Validation (Mar-Apr): ARIMA: OPEN									
Time	Year	Month	Actual Values	FOR_ARI	LOW_ARI	UPP_ARI	Actual in PI_ARI	MAE_ARI	MSE_ARI
311	2023	Mar	156.300	147.805	139.281	156.328	YES	8.495	72.165025
312	2023	Mar	156.740	147.525	135.727	159.322	YES	9.215	84.916225
313	2023	Mar	162.140	147.523	132.95	162.096	YES	14.617	213.656689
314	2023	Mar	165.000	148.683	131.198	166.168	YES	16.317	266.244489
315	2023	Apr	166.840	149.234	129.566	168.903	YES	17.606	309.971236
316	2023	Apr	166.320	149.247	127.842	170.651	YES	17.073	291.487329
317	2023	Apr	168.160	149.122	126.132	172.111	YES	19.038	362.445444
318	2023	Apr	169.850	149.849	125.072	174.625	YES	20.001	400.040001
319	2023	Apr	169.850	150.529	124.113	176.945	YES	19.321	373.301041

As anticipated, the regression model has a worst prediction accuracy of 44%, primarily due to its narrow prediction bounds. The ARIMA and double exponential smoothing model has a better prediction accuracy of 100%. Despite this, we chose the ARIMA model to forecast the values of Open, as it performed relatively better than the other two models relatively smaller MSE 18.90 as compared to double exponential smoothing model.

To summarize, the following models have been chosen to forecast the values of High, Low, and Open:
- High: Double Exponential Smoothing with level 0.8 and trend 0.1.
- Low: Double Exponential Smoothing with level 0.8 and trend 0.4.
- Open: ARIMA (3,1,3)

Testing

We have now reached the final step of our modeling process, which is testing. We will test the models for the values from March 2023 to April 2023 using two approaches.

- First, we will use the forecasted values generated by the optimal models for High, Low, and Open to forecast Close.
- Second, we will use the actual available values for High, Low, and Open from March- April 2023 to forecast Close.

Actual

Time	Year	Month	Actual Values	FOR_REG	LOW_REG	UPP_REG	Actual in PI_REG
311	2023	Mar	156.300	149.784	145.473	154.094	NO
312	2023	Mar	156.740	154.190	149.859	158.522	YES
313	2023	Mar	162.140	158.957	154.602	163.311	YES
314	2023	Mar	165.000	159.457	155.035	163.878	NO
315	2023	Apr	166.840	162.789	158.402	167.177	YES
316	2023	Apr	166.320	162.563	158.127	166.999	YES
317	2023	Apr	168.160	165.040	160.595	169.484	YES
318	2023	Apr	169.850	165.395	160.853	169.937	YES
319	2023	Apr	169.850	167.186	162.679	171.693	NO

Optimal

Time	Year	Month	Actual Values	FOR_REG	LOW_REG	UPP_REG	Actual in PI_REG
311	2023	Mar	156.300	141.437	137.322	145.552	NO
312	2023	Mar	156.740	141.285	137.140	145.429	NO
313	2023	Mar	162.140	141.361	137.189	145.533	NO
314	2023	Mar	165.000	142.395	138.214	146.576	NO
315	2023	Apr	166.840	142.925	138.722	147.128	NO
316	2023	Apr	166.320	143.009	138.774	147.244	NO
317	2023	Apr	168.160	142.977	138.704	147.249	NO
318	2023	Apr	169.850	143.648	139.350	147.946	NO
319	2023	Apr	169.850	144.278	139.950	148.606	N0

As anticipated, the optimal regression model performed poorly when provided with forecasted values for High, Low, and Open. It did not even have single actual value within the prediction bounds. However, when given actual values for High, Low, and Open, the regression model performed well with a prediction accuracy of 66.66%.

Conclusion:

To summarize, we can infer that the fitted regression model for Close remains the preferred model due to its low MSE and inclusion of the impact of High, Low, and Open in the model. However, during the testing phase, we encountered a drawback of the model as we can develop a more advanced model for the terms High, Low, and Open. It was noticed that the regression model performed exceptionally well when supplied with the actual values for these terms. Thus, there is a potential for building more effective prediction models for High, Low, and Open in the future.

Future scope:

- FFORMA, RNN, and LSTM models can be implemented for forecasting future values.
- Adding more features in dataset.
- Enhancement in the models of the fundamental data for Open, High, and Low.

Appendix:

Predictors	**Abbv.**
Time	T
Time*Time	T-T
Time*Time*Time	T-T-T
Time*Time*Time*Time	T-T-T_T
Time*Time*Time*Time*Time	T-T-T-T-T

Data preprocessing: Code:

```
import pandas as pd
df= pd.read_csv("/content/AAPL-4.csv")
df = df.dropna()
df['Date'] = pd.to_datetime(df['Date'])
df.to_csv('final_data.csv')
```

Optimal data for forecasting close

Time	Year	Month	Actual Values	High	Low	Open
311	2023	Mar	156.300	151.707	144.058	147.805
312	2023	Mar	156.740	151.954	144.42	147.525
313	2023	Mar	162.140	152.202	144.783	147.523
314	2023	Mar	165.000	152.449	145.145	148.683
315	2023	Apr	166.840	152.696	145.507	149.234
316	2023	Apr	166.320	152.943	145.869	149.247
317	2023	Apr	168.160	153.191	146.231	149.122
318	2023	Apr	169.850	153.438	146.594	149.849
319	2023	Apr	169.850	153.685	146.956	150.529

References:

https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analysis-with-arima-model/

https://finance.yahoo.com/quote/AAPL/history/

https://towardsdatascience.com/stock-market-anomalies-and-stock-market-anomaly-detection-are-two-different-things

https://ieeexplore-ieee-org.ezproxy.rit.edu/document/9776372

https://ieeexplore-ieee-org.ezproxy.rit.edu/document/9765177