Stat-773

Project Report

Time Series Prediction and Forecasting of NASDAQ-100 Index

BY:

Addl Tariq | at1816@rit.edu

Prasanna Rodrigues | pr1408@rit.edu

Introduction:

The Nasdaq Composite Index is a market capitalization-weighted index of more than 3,700 stocks listed on the Nasdaq stock exchange. As a broad index heavily weighted toward the important technology sector, the Nasdaq Composite Index has become a staple of financial markets reports. As a group of international graduate students, we are really enticed to invest our savings in the stock market. Given, the minimal to not existing experience in stock invest portfolio, we decided to gauge and evaluate the behavior of the NASDAQ market, before making any investments. For this purpose, we gathered a data for the tech-heavy NASDAQ-100 index for the past 5 years from 2017-2022 to forecast the potential time series and looking for the market investment scope.

Data Collection:

We gathered datasets from online websites and google datasets for daily NASDAQ-100 index, Covid data, and international Brent Oil information. We also gathered information for monthly US unemployment rate to be added to the formation of our dataset. The NASDAQ-100 index file consisted of High, Low, Open, Volume and Close price (our response for the time series) for the index stock on a daily level.

	Date	Open	High	Low	Close	Adj Close	Volume
2097	2022-09-29	11334.57031	11339.91016	11038.94043	11164.78027	11164.78027	4.516630e+09
2098	2022-09-30	11123.11035	11296.16016	10966.95020	10971.21973	10971.21973	4.649710e+09
2099	2022-10-01	NaN	NaN	NaN	NaN	NaN	NaN

The Unemployment rate was available to us from 1948 to present day hence we had to filter out the data from 2017-2022 as shown below. The Brent oil dataset was also available on a daily scale.

	Unemp_rate	Month_year
0	3.4	1948-01
1	3.8	1948-02
2	4.0	1948-03

	Date	COB_price
0	2012-11-06	109.27
1	2012-11-07	108.21
2	2012-11-08	107.23
3	2012-11-09	108.61

The Covid-19 data was available from March 2019 to 2022 on a daily interval. The dataset contained features related to total cases, new cases, total deaths, ICU patients, hospitalized patients, positive rate and total vaccinations and stringency index (official social distancing regulations impact index).

	Date	total_cases	new_cases	total_deaths	new_deaths	icu_patients	hosp_patients	positive_rate	tests_per_case	total_vaccinations	people_vaccinate
219539	2022-10-31	97503019.0	43008.0	1070508.0	123.0	2567.0	21762.0	NaN	NaN	640878616.0	266397317.
219540	2022-11-01	97550350.0	47331.0	1070907.0	399.0	NaN	NaN	NaN	NaN	640913400.0	266401911.
219541	2022-11-02	97622888.0	72538.0	1071649.0	742.0	NaN	NaN	NaN	NaN	NaN	Nal
219542	2022-11-03	97692050.0	69162.0	1072222.0	573.0	NaN	NaN	NaN	NaN	NaN	Nal
219543	2022-11-04	97729653.0	37603.0	1072561.0	339.0	NaN	NaN	NaN	NaN	NaN	Nal

Data Pre-processing:

We used the Python programming language and its libraries to consolidate and combine the datasets for our required time frame from 2017-2022 (Python script is provided in submission). The combined daily dataset had one specific issue of missing values for almost all the features. This is due to non-captured values in daily dataset for weekend and national holidays within US. Secondly, the pandemic came in 2019 hence we do not have values for the prior years.

```
1 date_i_ndx_unemp_bo_cv.isnull().sum().sort_values(ascending=False)
```

```
people_fully_vaccinated    1442
people_vaccinated          1442
total_vaccinations         1442
hosp_patients              1291
icu_patients               1291
tests_per_case             1266
positive_rate              1266
new_deaths                 1156
total_deaths               1154
new_cases                  1117
stringency_index           1116
total_cases                1116
Volume                      653
Adj Close                   653
Close                       653
Low                         653
High                        653
Open                        653
COB_price                   638
Month_year                    0
Unemp_rate                    0
Date                          0
dtype: int64
```

To solve this issue to form a continuous time series for our forecasting of NASDAQ-100 index, we applied Linear Interpolation to fill out missing values for all the features such as COB price (oil), Open, High, Low, Close, Volume, and Covid-19 features. We also replaced null values to zeros to from 2017 – Feb 2019 for Covid Features to ensure reliability of values.

```
1 date_i_ndx_unemp_bo_cv['Close']= date_i_ndx_unemp_bo_cv['Close'].interpolate(method='linear')
```

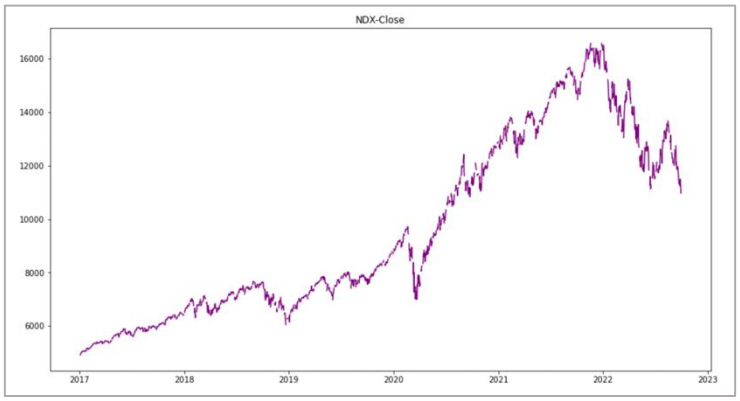


Figure 1 Y-response (NASDAQ-100 Close index) with missing values

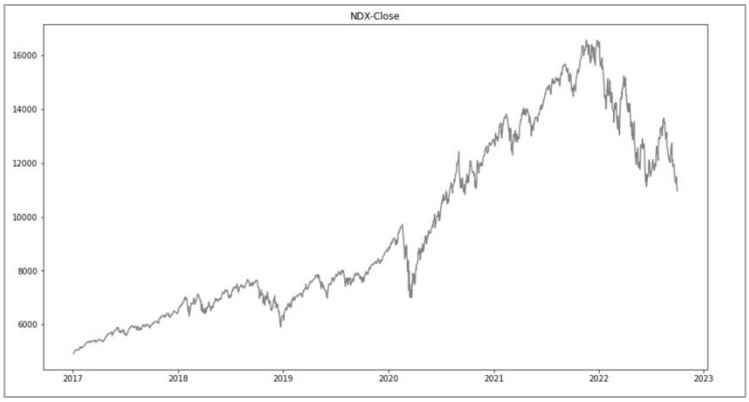


Figure 2 Y-response (NASDAQ-100 Close index) after Linear Interpolation

We used the daily dataset for modeling and realized the extreme issue of outliers impacting our models with dropping the Durbin Watson score below 1 for regression techniques. Hence, to cater this issue we aggregate the data to a weekly level, by taking the starting of week price as Open, ending price of the week as Close, with maximum peak during the week as High and the minimum peak during the week as

Low. We also average the price of the oil and summed values for new covid-19 cases (NC). We took the weekly mean for the Stringency index (SI).

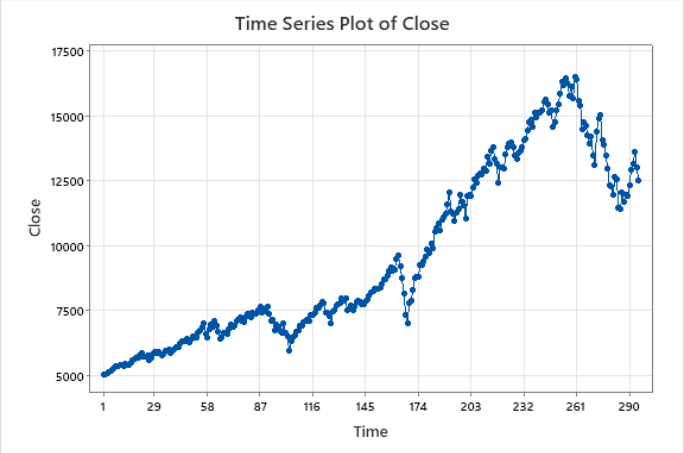
```
gg = dd.resample('W').agg({'Open':'first', 'High':'max', 'Low':'min', 'Adj Close':'last', 'Close':'last', 'total_ca
```

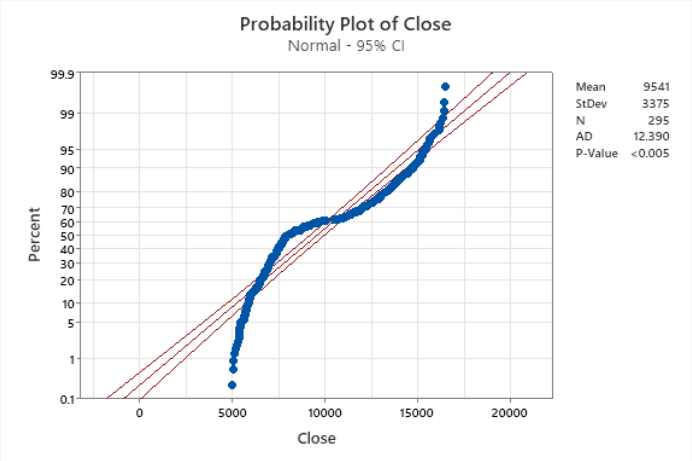
The weekly aggregation data for was compiled to be form a CSV for the Minitab file for modeling.

Date	Open	High	Low	Adj Close	Close	total_cases	Unemp_rate	new_cases	stringency_index
2022-09-04	12488.45996	12594.33984	12012.98047	12098.44043	12098.44043	94760242.0	3.5	550600.0	24.225714
2022-09-11	12106.57031	12610.42969	11928.80957	12588.29004	12588.29004	95257606.0	3.5	497364.0	25.495714
2022-09-18	12649.23047	12752.83008	11710.25977	11861.37988	11861.37988	95664107.0	3.5	406501.0	25.988571
2022-09-25	11753.59961	12062.51953	11169.66016	11311.24023	11311.24023	96068353.0	3.5	404246.0	25.968571
2022-10-02	11283.13965	11546.87012	10966.95020	10971.21973	10971.21973	96387457.0	3.7	319104.0	25.951667

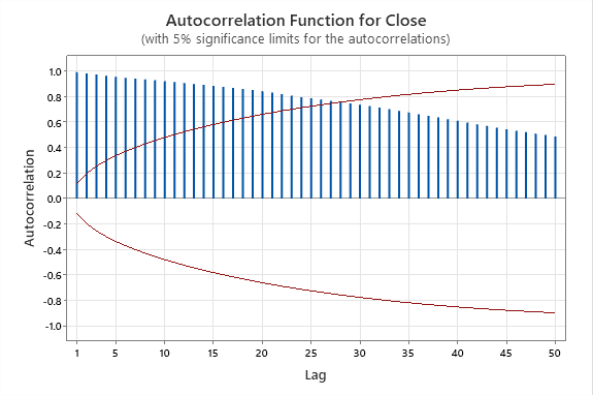
Training:

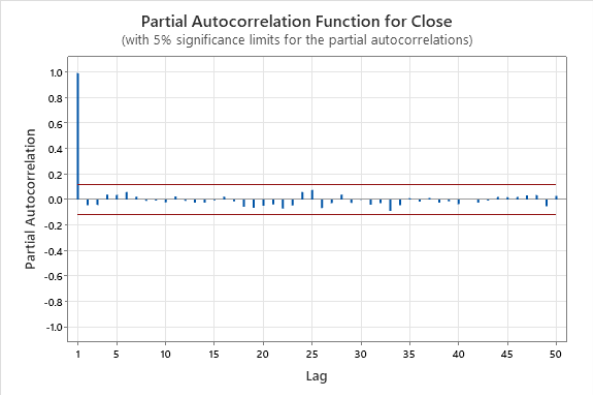
We start by analyzing the time series plot of the Nasdaq closing index. The time series plot indicates that the series has a mixture of trends. The series seems to follow an increasing trend till the time 261 and then starts decreasing. The probability plot of the series indicates that data doesn't follow the normality as the p value is less than 0.05.





To identify if the series has some seasonal behavior, we plot the ACF and PACF for the series. We find the ACF for the series and observe that many of the lags lie above the 95% confidence interval concluding that the series is not stationary. The ACF plot trend doesn't seem to be seasonal as verified by the PACF plot, hence we can conclude that the behavior of the series is not seasonal and is cyclic in nature. The PACF shows Lag 1 passes the 95% confidence interval.





To find the optimal model we decided to divide the series into three parts. The first part was the training data that contained values from Jan – 2017 to May – 2017. The second part was the validation data that contained values from June - 2022 to August – 2022. And the final part was testing data that contained values from September – 2022 to November – 2022.

To model this series, we consider multiple approaches such as the following:

- Regression Modeling
- Double Exponential Smoothing
- ARIMA Modeling

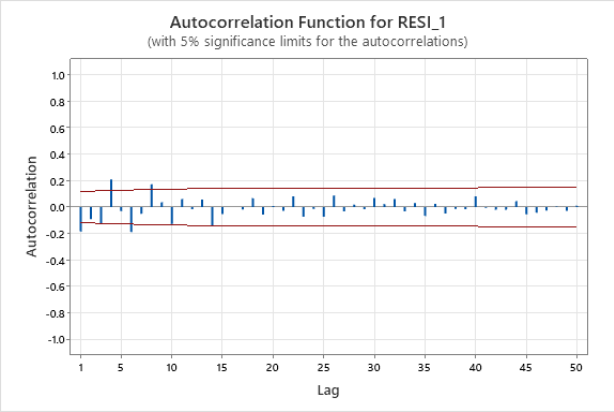
Dummy variables methodology, trigonometric models, Winter Holts method, and SARIMA were not considered due to the non-seasonal behavior of the series. The single exponential smoothing handles no trend, and non-seasonal series hence cannot be applicable here.

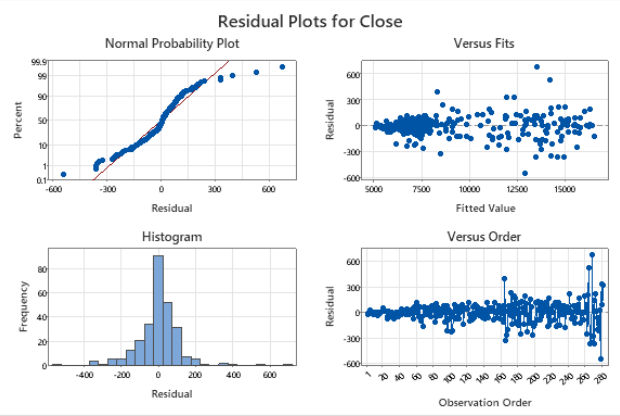
The Time series plot indicated that the series has some polynomial trend to it. Hence, we fit a regression model that included polynomial terms of Time. We went on adding higher degree polynomial terms of Time in the model and stopped when it was no longer significant. Since the series is not monthly seasonal, we won't use the variable month for the regression modeling of this series. The first regression model that we fitted to the series included the terms Time, Open, High, and Low. The Durbin-Watson score of this model is relatively high at 2.3437. The R-square value is impactful at 99.85%. Also, there is not much difference between the R square, R square adjusted and R square predicted value. We also checked the time cube interaction which was significant but didn't change the Durbin-Watson score significantly. We also wanted to see if Covid-19, US unemployment rate and Brent oil had any statistically significant effect on the market, so we even included those terms in the model, but they all turned out to be statistically non-significant.

Regression model								
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	18,000	134.164	99.85%	99.84%	99.84%	2.3437
2	T, T-T, T-T-T, O, H, L	O, H, L	17,902	133.798	99.85%	99.84%	99.83%	2.35805
3	O, H, L, TC, UN, NC, SI	O, H, L	17,788	133.371	99.85%	99.85%	99.83%	2.382

4	T, O, H, L, TC, UN, NC, SI	O, H, L	17,853	133.614	99.85%	99.84%	99.83%	2.382
5	T, T-T, T-T-T, T-T-T-T, T-T-T-T-T, O, H, L, TC, UN, NC, SI	T-T-T-T, T-T-T-T-T, O, H, L	17,166	131.02	99.86%	99.85%	99.83%	2.396
6	T, T-T, T-T-T, T-T-T-T, T-T-T-T-T, O, H, L	T-T-T-T, T-T-T-T-T, O, H, L	16,993	130.357	99.86%	99.85%	99.84%	2.391
7	**T, T-T, T-T-T, T-T-T-T, T-T-T-T-T, O, H, L, Lag1**	**T-T-T-T-T, O, H, L, Lag1**	**15,632**	**125.029**	**99.87%**	**99.86%**	**99.85%**	**2.3429**
8	T, T-T, T-T-T, T-T-T-T, T-T-T-T-T	T-T, T-T-T, T-T-T-T, T-T-T-T-T	206,444	454.361	98.24%	98.21%	98.16%%	0.3935

We wanted our model to have better precision to achieve that we need to obtain a model with low mean square error and high prediction accuracy. To decrease the mean square error and correlation between the residuals of the model higher degree polynomial terms were also added to the model, and it was found that time up to degree 5 was significant. Since we had very high correlation at Lag 1 in the PACF plot we added Lag 1 of close to the model to reduce the autocorrelation between residuals. The ACF of residual for this regression model indicates that there isn't any extreme correlation within the residual since except for a few lags all the other lags are within the 95% confidence intervals.



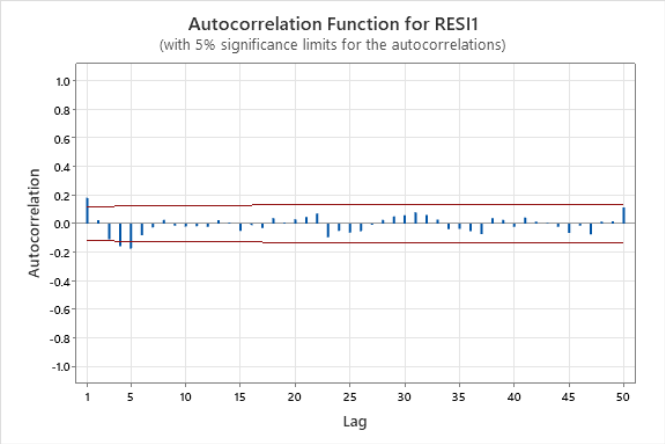


We now move on to our second modeling approach which is Double Exponential Smoothing. We use Level and Trend levers of the technique to fit the best model. We initially set the level and the trend at 0.2 to observe the bad MSD score of 218688 which doesn't demonstrate a good model compared to our regression model baseline. We first tried to decrease the Trend by keeping the Level constant to observe the change in MSD decreasing normally. Then we changed the trend to smaller values to review the decreasing MSE. Our best double exponential smoothing model is **6** from the table, with a Trend of 0.1 and a Level of 0.8. We observed that the change in level or trend further didn't make a substantial impact on the MAD, MAPE, and MAD metrics, hence we stopped the process.

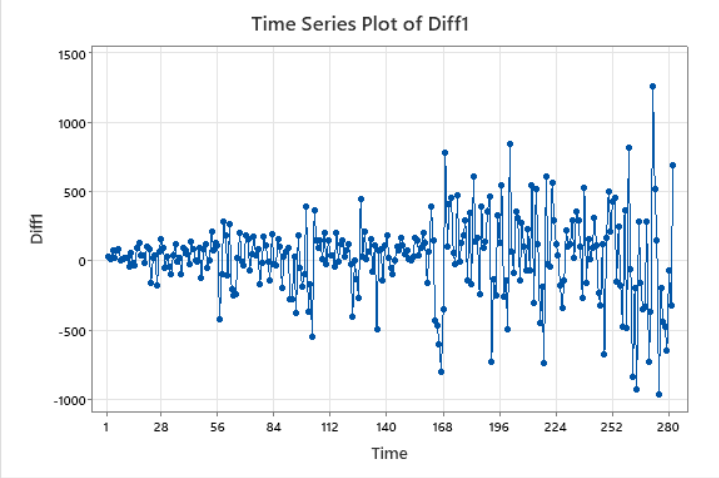
Double Exponential Smoothing

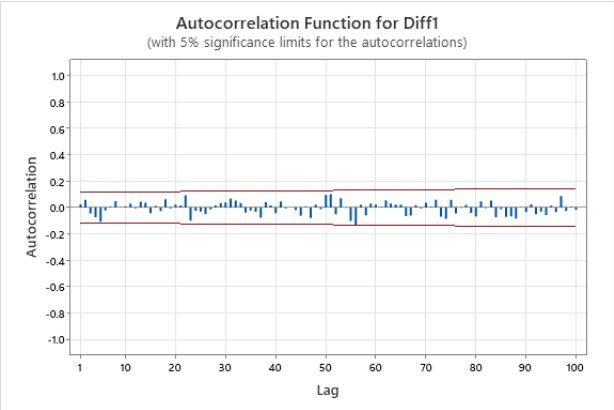
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	4	332	218,688
2	0.2	0.1	4	332	208,298
3	0.2	0.05	4	333	214,537
4	0.3	0.1	3	285	161,685
5	0.5	0.1	3	242	118,550
6	**0.8**	**0.1**	**2.2**	**211.3**	**94,019.2**

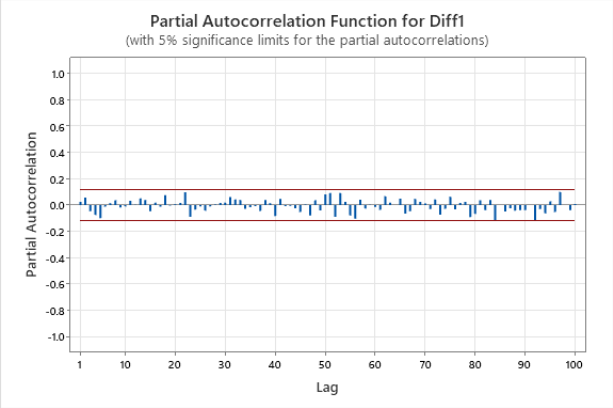
We now observe the ACF of **model 6** from this approach, which demonstrates that except for few lags other all are with the 95% CI which represents that double exponential smoothing model residuals do not have extreme autocorrelation. But since we want our precision to be high and MSE of 94019 is significantly worse compared to our regression model. Hence, we **discard this approach**.



Now we move to the ARIMA modeling technique. We first find the Difference of Lag 1 from the time series to remove the trend for the series. The plot of Diff1 below demonstrates that the Diff1 series can potentially be white noise. We found the ACF and PACF of Diff1 to confirm the hypothesis, where we observe that the series now indeed is a white noise.







Now observing the above ACF and PACF plots we did not have any lags outside the 95% confidence interval. So, we tried different combinations of AR and MA for the ARIMA model. To further check if any model results will satisfy the series, we checked other MA values keeping the AR term constant.

ARIMA

Sr.	AR	Difference	MA	Significant Terms	MSE	Goodness of Fit	SS
1	1	1	1	-	83,284.0	PASS	23,153,216
2	1	1	2	-	83,525.4	PASS	23,136,539
3	1	1	3	-	82,924.8	PASS	22,887,252
4	1	1	5	-	82,696.0	PASS	22,658,697
5	4	1	5	AR3, MA3	83,151.1	PASS	22,533,961
6	3	1	3	AR1, AR2, MA1, MA2	82,650.5	PASS	22,646,235

Like our Double exponential smoothing approach, the ARIMA models also have significantly high MSE values compared to the regression model. So, we discarded this model.

Validation:

Our best model to predict the NASDAQ index was the regression model as both Double Exponential Smoothing and ARIMA techniques were discarded based on having high MSE. Below we have the predicted values using our finalized regression model and the 7th ARIMA model to see the difference between the precision values for both these models

Validation (June-Aug)									
Time	Year	Month	Actual Values	FOR_ARI	LOW_ARI	UPP_ARI	Actual in PI_ARI	MAE_ARI	MSE_ARI
283	2022	June	12,582.43001	12,686.0	12,123.3	13,248.7	YES	103.6	107,27.3
284	2022	June	11,469.82031	12,818.2	12,014.0	13,622.4	NO	1348.4	1,818,151.6
285	2022	June	11,406.375	12,917.4	11,904.2	13,930.6	NO	1511.1	2,283,276.2
286	2022	June	12,040.77669	12,902.3	11,707.4	14,097.2	YES	861.6	742,274.9
287	2022	July	11682.79004	12840.3	11499.1	14,181.5	YES	1,157.5	1,339,837.5

288	2022	July	11948.75032	12845.4	11384.9	14,306.0	YES	896.7	804,059.9
289	2022	July	11912.87337	12935.0	11364.0	14,506.1	YES	1,022.2	1,044,826.1
290	2022	July	12351.09668	13022.6	11338.8	14,706.4	YES	671.5	450,920.7
291	2022	July	12943.17676	13036.4	11241.1	14,831.7	YES	93.2	8,687.8
292	2022	August	13175.33692	13003.4	11106.9	14,899.8	YES	172.0	29,576.3
293	2022	August	13633.40983	13002.8	11017.1	14,988.5	YES	630.6	397,650.1
294	2022	August	13007.99349	13063.8	10994.0	15,133.6	YES	55.8	3,113.8
295	2022	August	12524.60351	13137.4	10982.3	15,292.5	YES	612.8	375,521.2

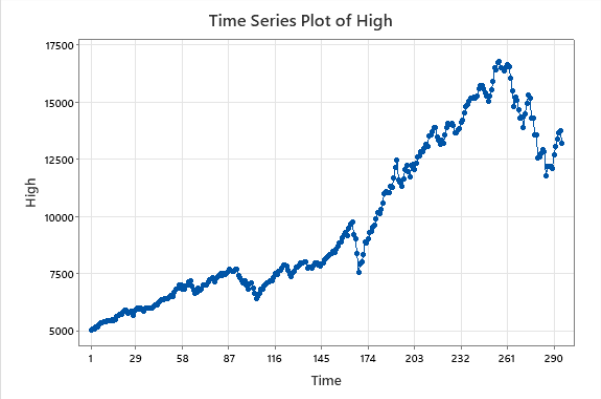
Validation (June-Aug)									
Time	Year	Month	Actual Values	FORE_REG	LOW_REG	UPP_REG	Actual in PI_REG	MAE_REG	MSE_REG
283	2022	June	12,582.43001	12,649.0	12,381.5	12,916.5	YES	66.6	4,432.1
284	2022	June	11,469.82031	11,717.2	11,445.7	11,988.8	YES	247.4	61,220.2
285	2022	June	1,1406.375	11,335.4	11,065.1	11,605.8	YES	71.0	5,035.7
286	2022	June	12,040.77669	11,991.4	11,716.4	12,266.3	YES	49.4	2,440.2
287	2022	July	11,682.79004	11,434.5	11,156.0	11,713.0	YES	248.3	61,650.8
288	2022	July	11,948.75032	11,816.2	11,535.6	12,096.8	YES	132.6	17,569.8
289	2022	July	11,912.87337	11,546.1	11,258.3	11,833.9	NO	366.8	134,514.6
290	2022	July	12,351.09668	12,303.4	12,003.2	12,603.6	YES	47.7	2,272.7
291	2022	July	12,943.17676	12,492.4	12,187.5	12,797.2	NO	450.8	203,212.1
292	2022	August	13,175.33692	12,939.1	12,620.4	13,257.9	YES	236.2	55,791.3
293	2022	August	13,633.40983	13,069.0	12,736.6	13,401.4	NO	564.4	318,516.6
294	2022	August	13,007.99349	12,817.1	12,475.0	13,159.1	YES	190.9	36,457.8
295	2022	August	12,524.60351	12,352.9	12,005.0	12,700.9	YES	171.7	29,471.0

The ARIMA model had a prediction accuracy of 84.62% with the Mean Absolute Error of 702.8 and Root Mean Square Error of 846.2. Whereas the regression model had a prediction accuracy of 76.92% with the Mean Absolute Error of 218.7 and Root Mean Square Error of 267.8. In terms of precision regression model did way better than the ARIMA model and had an acceptable prediction accuracy. So, we finalized the regression model as our best model.

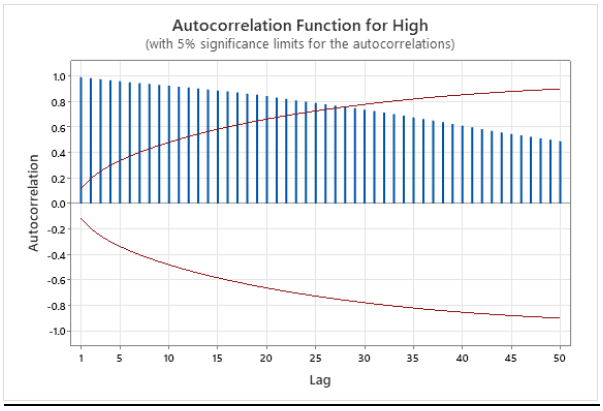
Building underlined models:

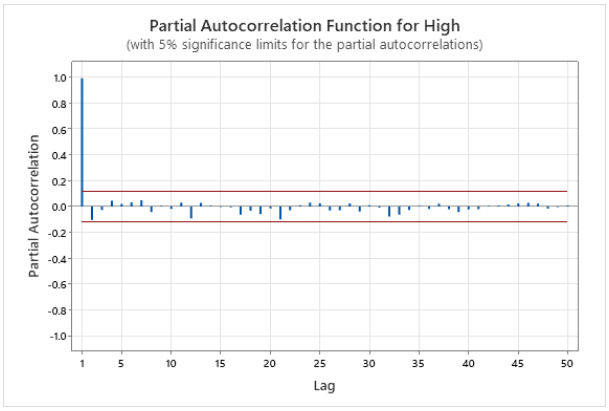
We have trained and validated our model the next step now is to test it. Since our best model included the open, high, and close terms we need to find optimal models for each of those terms. We are going to train all these models on the data from January – 2017 to August- 2022 and validate them of the data from September – 2022 to November – 2022. Open might have some dependencies on high and low so we start by the series for high.

We start by analyzing the time series plot of High. The time series plot for high is like close and indicates that the series has a mixture of trends. Even this series seems to follow an increasing trend till the time 261 and then starts decreasing.



Looking at the ACF for the series we observe that many of the lags lie above the 95% confidence interval concluding that the series is not stationary. The ACF plot trend doesn't seem to be seasonal as verified by the PACF plot, hence we can conclude that the behavior of the series is not seasonal and is cyclic in nature. The PACF shows Lag 1 passes the 95% confidence interval.





As this series is like the series of close, we use the same three models to model this series. The Time series plot indicated that the series has some polynomial trend to it. Hence, we fit a regression model that included polynomial terms of Time. We went on adding higher degree polynomial terms of Time in the model and stopped when it was no longer significant. Since we had very high correlation at Lag 1 in the PACF plot we added Lag 1 of open to the model to reduce the autocorrelation between residuals. Even after adding lag 1 there was still significantly high autocorrelation between the residuals, so we decided to add lag 2 as well. The Durbin-Watson score of this model is pretty good at 2.0536. The R-square value is impactful at 99.59%. The final regression model is given below:

High - Regression model - Training (Jan-Aug)

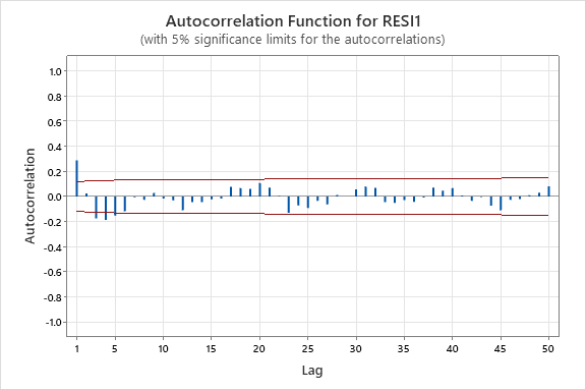
Sr.	Terms Included	Significant Terms	MSE	S	R-square	R-Square Adjusted	R Square Predicted	Durbin Watson
1	**T, T-T, T-T-T, T-T-T-T, Lag1, Lag2**	**T, T-T, T-T-T, T-T-T-T, Lag1, Lag2**	**49,604**	**222.718**	**99.59%**	**99.58%**	**99.56%**	**2.0536**

We now move on to our second modeling approach which is Double Exponential Smoothing. We use the same Level and Trend levers of the technique as before to fit the best model.

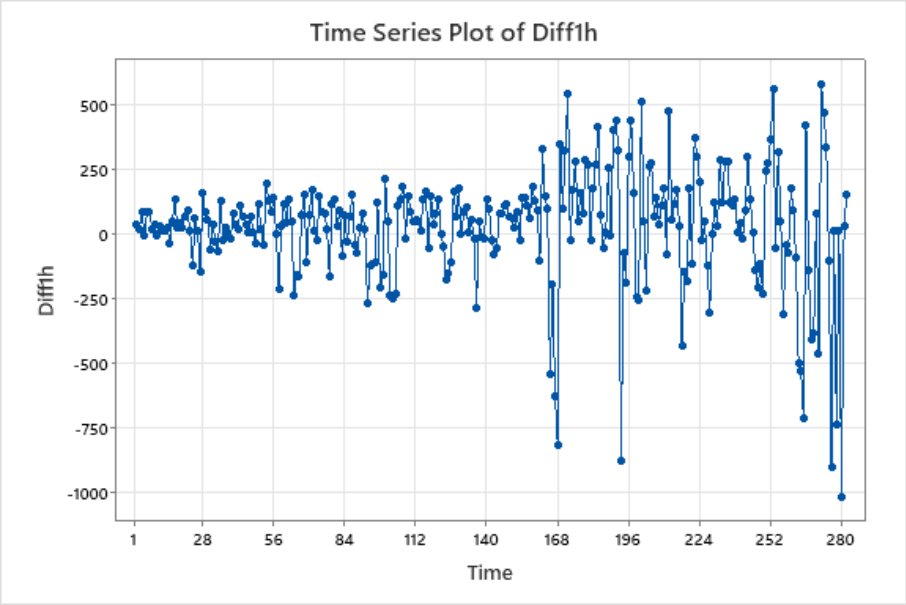
High Double Exponential Smoothing

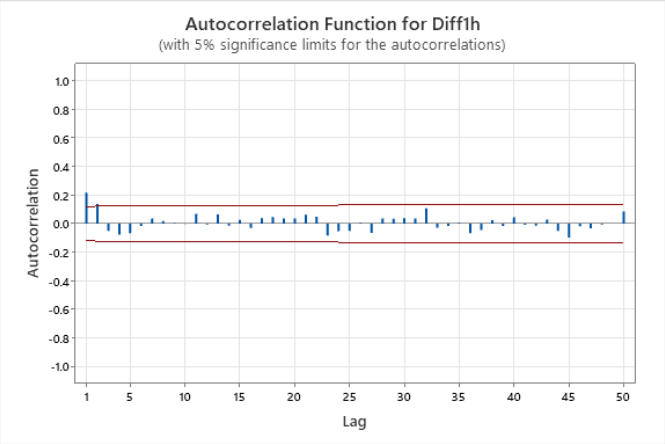
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	3	311	198,189
2	0.4	0.4	2	227	114,295
3	0.8	0.2	1.7	172.6	66,168

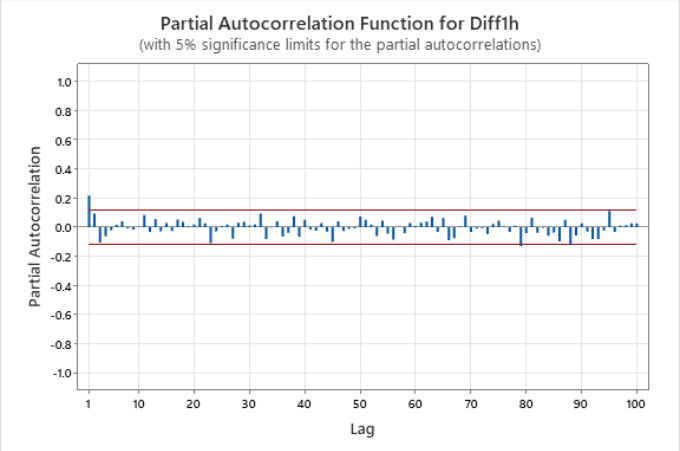
We now observe the ACF of **model 3** from this approach, which demonstrates that except for few lags other all are with the 95% CI which represents that double exponential smoothing model residuals do not have extreme autocorrelation.



Now we move to the ARIMA modeling technique. We first find the Difference of Lag 1 from the time series to remove the trend for the series. The plot of Diff1 below demonstrates that the Diff1 series can potentially be white noise. We found the ACF and PACF of Diff1 to confirm the hypothesis, where we observe that the series does have few lags above 95% CI.





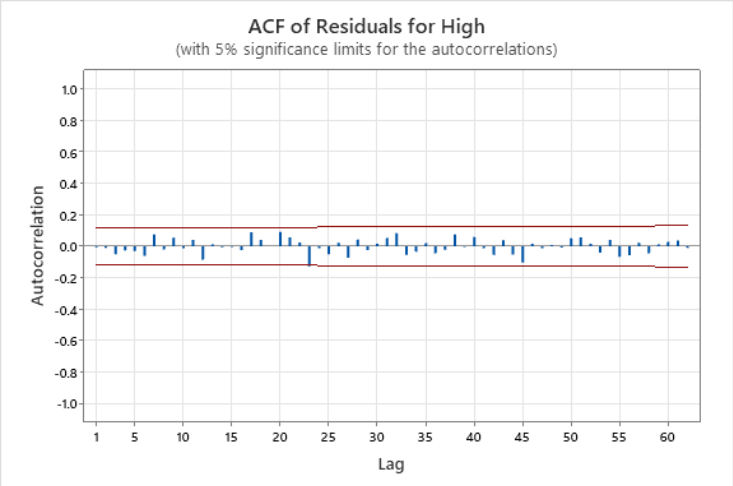


Now observing the above ACF and PACF plots we did have lag 1 outside the 95% confidence interval for both the plots and even Lag 2 for ACF plots seems to touch the 95% CI. So, the possible combinations of ARIMA are (1,1,1), (1,1,2), (2,1,1) and (2,2,2). To get the optimal ARIMA we used forecast with the optimal ARIMA option from Minitab.

High - ARIMA

Sr.	AR	Difference	MA	Significant Terms	MSE	Goodness of Fit	SS
1	**0**	**1**	**2**	**MA1, MA2**	**52144**	**PASS**	**15,173,915**

We now observe the ACF of the model from this approach, which demonstrates that all the lags are with the 95% CI which represents that ARIMA model residuals do not have extreme autocorrelation.



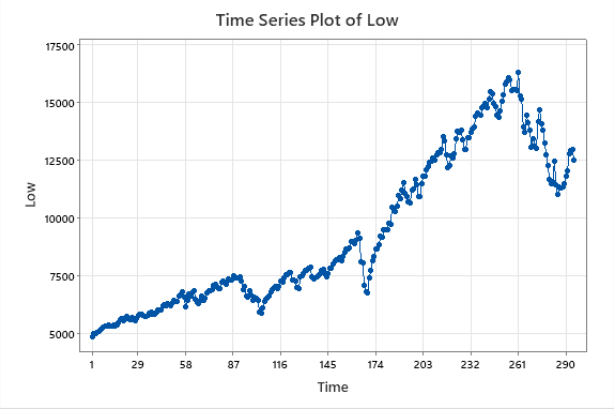
Since the MSE from all the three approaches was not significantly high compared to others we calculated the forecasted values from September – 2022 to November – 2022 for all the three approaches to validate the models.

High - Validation (Sep-Nov)

Time	Year	Month	Weekly Dates	Actual Values	Actual in PI_Reg	Actual in PI_ARI	Actual in PI_DE
296	2022	Sep	9/4/2022	12,594.34	Yes	No	No
297	2022	Sep	9/11/2022	12,705.36	Yes	Yes	Yes
298	2022	Sep	9/18/2022	12,752.83	No	Yes	Yes
299	2022	Sep	9/25/2022	12,062.52	Yes	Yes	Yes
300	2022	Oct	10/2/2022	11,546.87	Yes	No	No
301	2022	Oct	10/09/2022	11,660.55	Yes	No	Yes
302	2022	Oct	10/16/2022	11,152.89	Yes	No	No
303	2022	Oct	10/23/2022	11,421.38	Yes	No	Yes
304	2022	Oct	10/30/2022	11,681.85	No	Yes	Yes
305	2022	Nov	11/06/2022	11,574.39	No	Yes	Yes
306	2022	Nov	11/13/2022	11,855.9	No	Yes	Yes
307	2022	Nov	11/20/2022	12,024.95	No	Yes	Yes
308	2022	Nov	11/27/2022	11,866.42	No	Yes	Yes

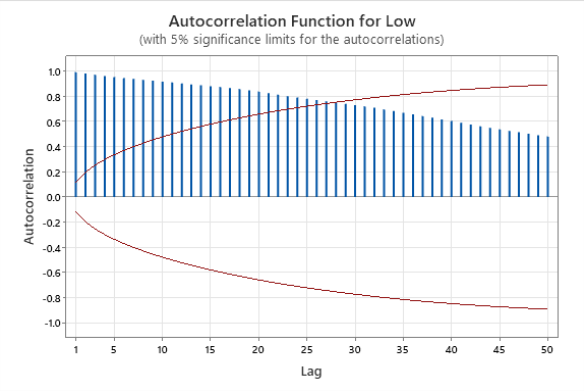
The regression model has a prediction accuracy of 53.85%, ARIMA has a prediction accuracy of 61.54% and Double Exponential Smoothing has a prediction accuracy of 76.92%. Even though Double Exponential Smoothing has better prediction accuracy it had the widest prediction bounds and hence had the least precision. ARIMA did not have great prediction accuracy but still had narrower predictions bounds and hence had more precision over Double Exponential Smoothing Model. So, for high, we decided to use the forecasted values from ARIMA model for testing.

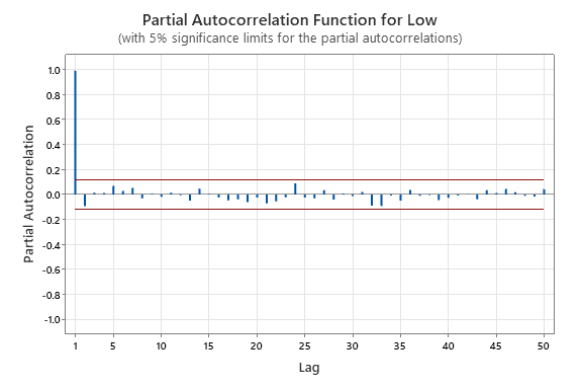
Now let's move on to low. The time series plot for low is also like close and indicates that the series has a mixture of trends. Even this series seems to follow an increasing trend till the time 261 and then starts decreasing.



Looking at the ACF for the series we observe that many of the lags lie above the 95% confidence interval concluding that the series is not stationary. The ACF plot trend doesn't seem to be seasonal as verified by

the PACF plot, hence we can conclude that the behavior of the series is not seasonal and is cyclic in nature. The PACF shows Lag 1 passes the 95% confidence interval.





As this series is like the series of close, we use the same three models to model this series. The Time series plot indicated that the series has some polynomial trend to it. Hence, we fit a regression model that included polynomial terms of Time. We went on adding higher degree polynomial terms of Time in the model and stopped when it was no longer significant. Since we had very high correlation at Lag 1 in the PACF plot we added Lag 1 of open to the model to reduce the autocorrelation between residuals. Even after adding lag 1 there was still significantly high autocorrelation between the residuals, so we decided to add lag 2 as well. The Durbin-Watson score of this model is pretty good at 2.007. The R-square value is impactful at 99.28%. The final regression model is given below.

Low - Regression model - Training (Sep-Nov)

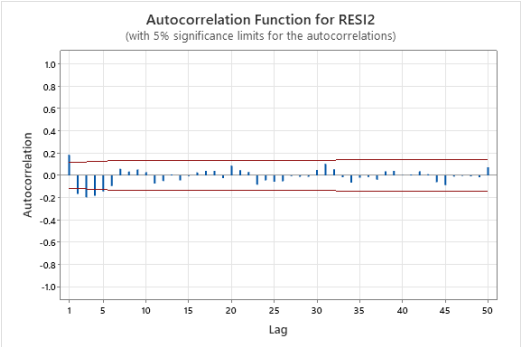
Sr.	Terms Included	Significant Terms	MSE	S	R-sq	R Square Adjusted	R Square Predicted	Durbin Watson
1	**T, T-T, T-T-T, T-T-T-T, Lag1, Lag2**	**T, T-T, T-T-T, T-T-T-T, Lag1, Lag2**	**78,344**	**279.901**	**99.28%**	**99.26%**	**99.23%**	**2.007**

We now move on to our second modeling approach which is Double Exponential Smoothing. We use the same Level and Trend levers of the technique as before to fit the best model.

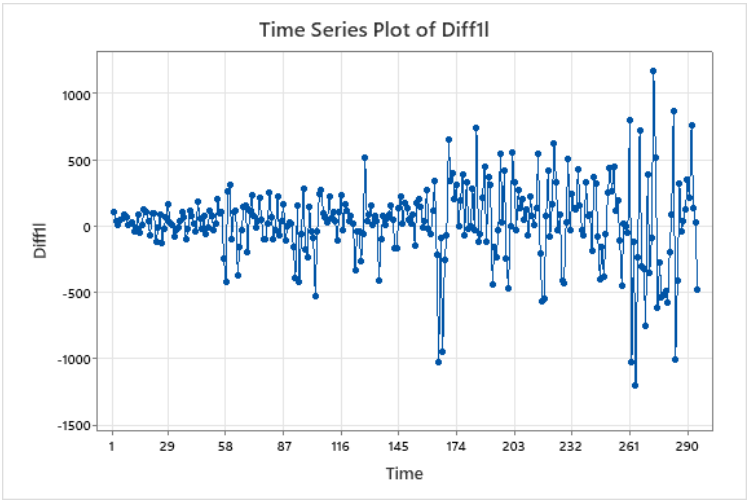
Low Double Exponential Smoothing

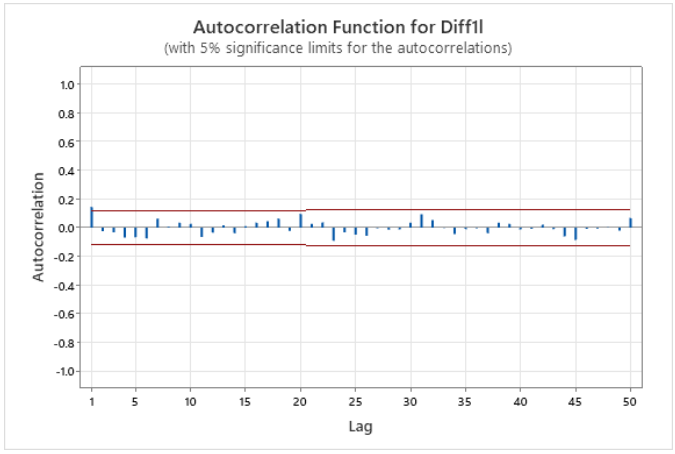
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.2	4	356	263,062
2	0.4	0.4	3	287	169,520
3	0.8	0.4	2	222	108,823

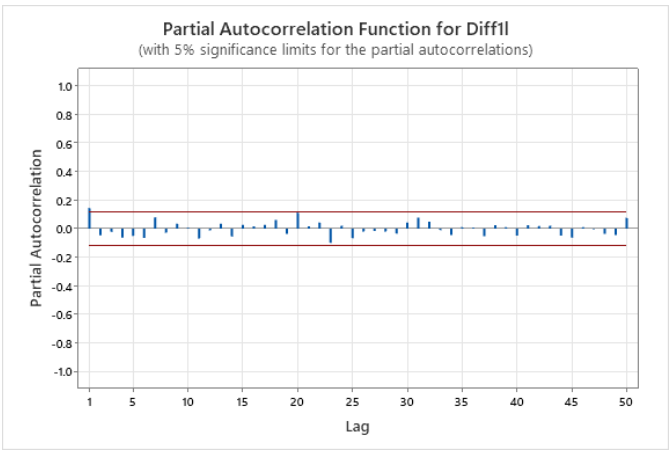
We now observe the ACF of **model 3** from this approach, which demonstrates that except for few lags other all are with the 95% CI which represents that double exponential smoothing model residuals do not have extreme autocorrelation.



Now we move to the ARIMA modeling technique. We first find the Difference of Lag 1 from the time series to remove the trend for the series. The plot of Diff1 below demonstrates that the Diff1 series can potentially be white noise. We found the ACF and PACF of Diff1 to confirm the hypothesis, where we observe that the series does have few lags above 95% CI.





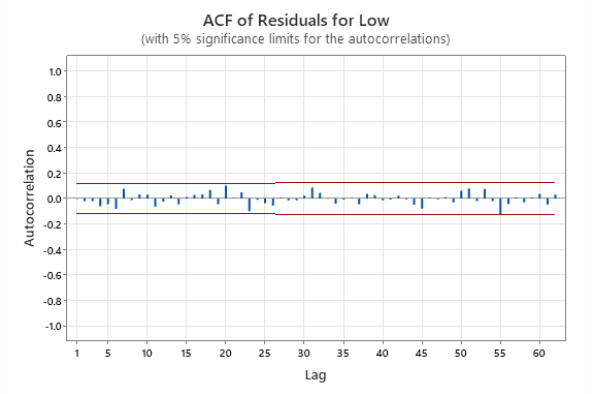


Now observing the above ACF and PACF plots we did have lag 1 outside the 95% confidence interval for both the plots. So, the possible combinations of ARIMA are (1,1,1). To get the optimal ARIMA we used forecast with the optimal ARIMA option from Minitab.

Low ARIMA - Low

SR No.	AR	Difference	MA	Significant Terms	MSE	Goodness of Fit	SS
1	**0**	**1**	**1**	**MA1**	**83117**	**PASS**	**24,270,174**

We now observe the ACF of the model from this approach, which demonstrates that all the lags are with the 95% CI which represents that ARIMA model residuals do not have extreme autocorrelation.



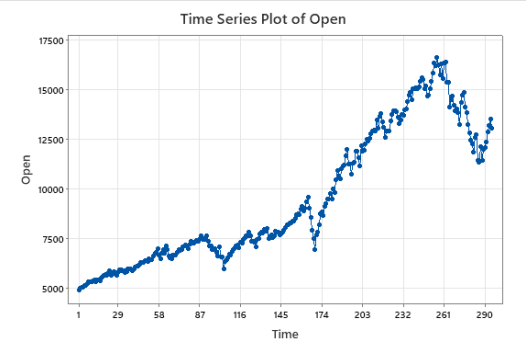
Since the MSE from ARIMA and regression approach was not significantly high compared to others we calculated the forecasted values from September – 2022 to November – 2022 for all the three approaches to validate the models.

Low - Validation (Sep-Nov)

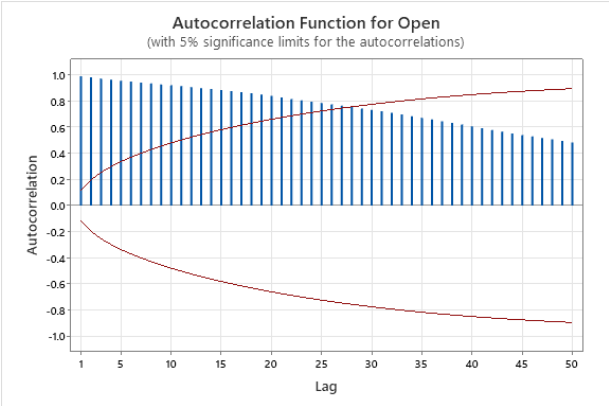
Time	Year	Month	Weekly Dates	Actual Values	Actual in PI_Reg	Actual in PI_ARI	Actual in PI_DE
296	2022	Sep	9/4/2022	11,982.17	Yes	Yes	No
297	2022	Sep	9/11/2022	11,928.81	Yes	Yes	Yes
298	2022	Sep	9/18/2022	11,710.26	Yes	Yes	Yes
299	2022	Sep	9/25/2022	11,169.66	Yes	No	No
300	2022	Oct	10/2/2022	10,966.95	Yes	No	No
301	2022	Oct	10/09/2022	10,880.05	Yes	No	No
302	2022	Oct	10/16/2022	10,440.64	Yes	No	No
303	2022	Oct	10/23/2022	10,959.74	No	Yes	Yes
304	2022	Oct	10/30/2022	11,166.49	No	Yes	Yes
305	2022	Nov	11/06/2022	10,632.39	No	Yes	Yes
306	2022	Nov	11/13/2022	10,790.35	No	Yes	Yes
307	2022	Nov	11/20/2022	11,519.38	No	Yes	Yes
308	2022	Nov	11/27/2022	11,503.34	No	Yes	Yes

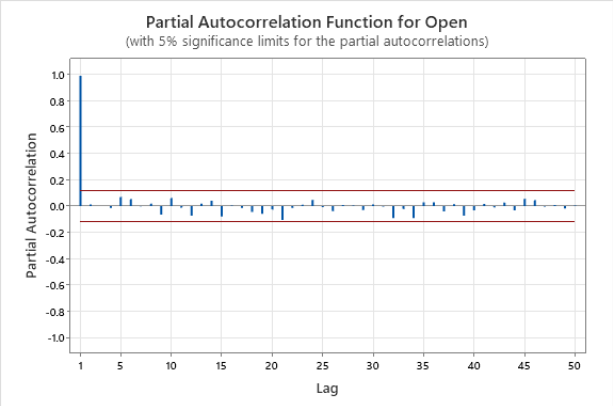
The regression model has a prediction accuracy of 53.85%, ARIMA has a prediction accuracy of 69.23% and Double Exponential Smoothing has a prediction accuracy of 61.53%. ARIMA did not have great prediction accuracy but still performed better over Double Exponential Smoothing and regression Model. So, for the series of Low we decided to use the forecasted values from ARIMA model for testing.

Now the final series we have to model is the series for Open. The time series plot for this series is also like close and indicates that the series has a mixture of trends. Even this series seems to follow an increasing trend till the time 261 and then starts decreasing.



Looking at the ACF for the series we observe that many of the lags lie above the 95% confidence interval concluding that the series is not stationary. The ACF plot trend doesn't seem to be seasonal as verified by the PACF plot, hence we can conclude that the behavior of the series is not seasonal and is cyclic in nature. The PACF shows Lag 1 passes the 95% confidence interval.





As this series is like the series of close, we use the same three models to model this series. The Time series plot indicated that the series has some polynomial trend to it. Hence, we fit a regression model that included polynomial terms of Time. We went on adding higher degree polynomial terms of Time in the model and stopped when it was no longer significant. Since Open mostly likely had dependency on high and low those terms were included in the model. Even though the time series seems to follow more than a linear trend none of the Time interaction terms were found significant and so the model was reduced to a point where it has just the Time term. The Durbin-Watson score of this model is high at 2.33 so there is some autocorrelation between residuals. The R-square value is impactful at 99.77%. The final regression model is given below.

Open - Regression model - Training (Sep-Nov)

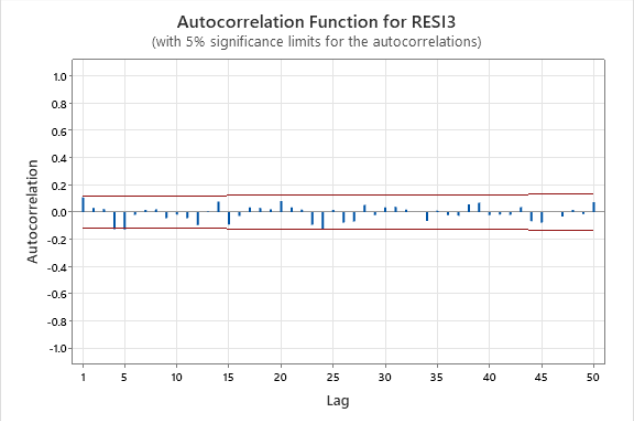
Sr.	Terms Included	Significant Terms	MSE	S	R-sq	R Square Adjusted	R Square Predicted	Durbin Watson
1	**T, H, L**	**H, L**	**26,974**	**164.237**	**99.77%**	**99.76%**	**99.75%**	**2.33**

We now move on to our second modeling approach which is Double Exponential Smoothing. We use the same Level and Trend levers of the technique as before to fit the best model.

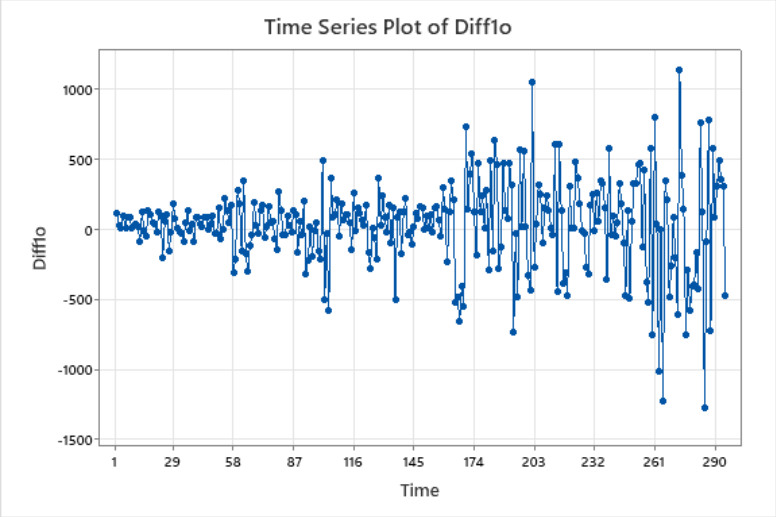
Open Double Exponential Smoothing

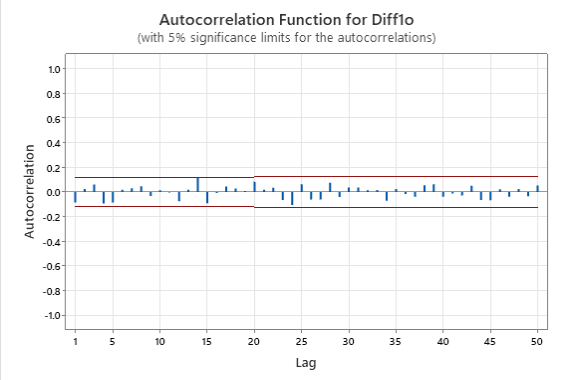
Sr.	Level	Trend	MAPE	MAD	MSD
1	0.2	0.4	4	345	250,830
2	0.4	0.4	3	276	154,655
3	0.8	0.4	2	236	118,092
4	0.8	0.2	2	232	108,920

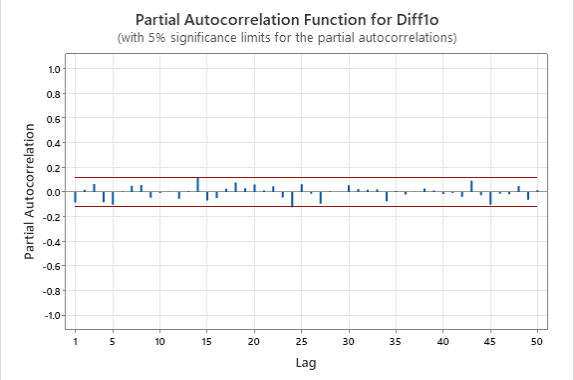
We now observe the ACF of **model 4** from this approach, which demonstrates that almost all lags are with the 95% CI which represents that double exponential smoothing model residuals do not have extreme autocorrelation.



Now we move to the ARIMA modeling technique. We first find the Difference of Lag 1 from the time series to remove the trend for the series. The plot of Diff1 below demonstrates that the Diff1 series can potentially be white noise. We found the ACF and PACF of Diff1 to confirm the hypothesis, where we observe that the series now indeed is a white noise.





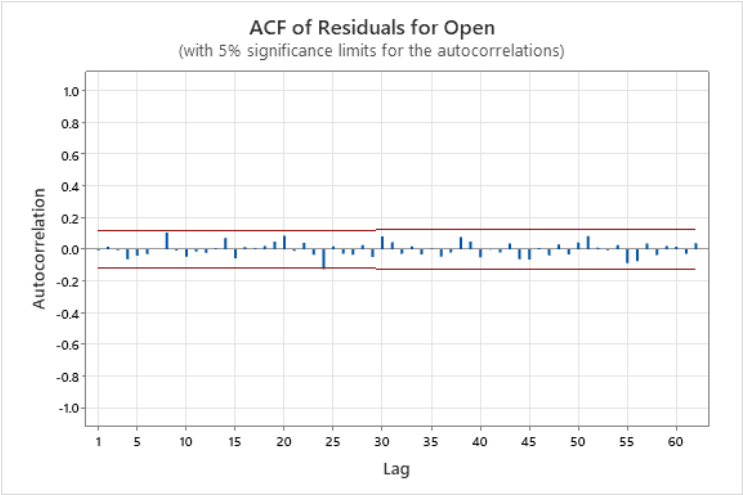


Now observing the above ACF and PACF plots we did have lag 1 and 2 outside the 95% confidence interval for both the plots. So, the possible combinations of ARIMA are (1,1,1), (1,1,2), (2,1,1) and (2,1,2). To get the optimal ARIMA we used forecast with the optimal ARIMA option from Minitab.

Open ARIMA

Sr.	AR	Difference	MA	Significant Terms	MSE	Goodness of Fit	SS
1	2	1	2	MA1, MA2, AR1, AR2	93,641	PASS	27,155,902

We now observe the ACF of the model from this approach, which demonstrates that all the lags are with the 95% CI which represents that ARIMA model residuals do not have extreme autocorrelation.



Since the MSE from regression was significantly smaller than the other approaches so regression will have better precision over others but at the same time it had more concise prediction bounds because of it might fail to capture the variability of the series. We calculated the forecasted values from September – 2022 to November – 2022 for all the three approaches to validate the models.

Open - Validation (Sep-Nov)

Time	Year	Month	Weekly Dates	Actual Values	Actual in PI_Reg	Actual in PI_ARI	Actual in PI_DE
296	2022	Sep	9/4/2022	12,488.46	Yes	Yes	No
297	2022	Sep	9/11/2022	12,178.04	No	No	No
298	2022	Sep	9/18/2022	12,649.23	Yes	Yes	Yes
299	2022	Sep	9/25/2022	11,753.6	No	No	No
300	2022	Oct	10/2/2022	11,283.14	No	No	No
301	2022	Oct	10/09/2022	11,059.17	No	No	No
302	2022	Oct	10/16/2022	11,048.51	No	No	No
303	2022	Oct	10/23/2022	10,967.25	No	No	No
304	2022	Oct	10/30/2022	11,321.11	No	No	Yes
305	2022	Nov	11/06/2022	11,465.21	No	Yes	Yes
306	2022	Nov	11/13/2022	10,900.83	No	No	Yes
307	2022	Nov	11/20/2022	11,728.11	No	Yes	Yes
308	2022	Nov	11/27/2022	11,623.35	No	Yes	Yes

As expected, the regression model has a very bad prediction accuracy of 15.38% because of having such concise prediction bounds, ARIMA has a prediction accuracy of 38.46% and Double Exponential Smoothing has a prediction accuracy of 46.15%. None of the models had prediction accuracy above 50% so realistically all these models are bad, and a better model is needed to model this series. But with our knowledge now, we couldn't find a better model for this series. So, we decided to go with the Double Exponential Smoothing model to forecast the values of Open as it worked relatively better over the other two models.

In summary, the final models selected to forecast the values of High, Low and Open are listed below.

- HIGH – ARIMA (0,1,2)
- LOW – ARIMA (0,1,1)
- OPEN – Double Exponential Smoothing with level 0.8 and trend 0.05.

Testing:

Now we are moving to the final step of our model that is testing. We are going to test the model for the values from September – 2022 to November – 2022. We decided to approach this in two ways:

- We are going to use the forecasted values using optimal model for High, Low and Open and then forecast close.
- We are going to use the actual available value from September – 2022 to November – 2022 for High, Low and Open and then forecast close.

The idea of applying these two approaches is to see how good of the model is the chosen regression model for close. Since all the underline models for High, Low and Open didn't have a good prediction accuracy using the forecasted values from this model is going to affect the forecast of close.

Close - Forecast (Sep-Nov)

Time	Year	Month	Weekly Dates	Actual Values	Actual in PI_Reg	Actual in PI_ARI	Actual in PI_Reg with Original values
296	2022	Sep	9/4/2022	12,054.88	No	Yes	Yes
297	2022	Sep	9/11/2022	12,689.24	Yes	Yes	Yes
298	2022	Sep	9/18/2022	11,922.64	No	Yes	Yes
299	2022	Sep	9/25/2022	11,273.15	No	Yes	Yes
300	2022	Oct	10/2/2022	11,143.56	No	No	Yes
301	2022	Oct	10/09/2022	10,964.47	No	No	No
302	2022	Oct	10/16/2022	10,939.04	No	No	Yes
303	2022	Oct	10/23/2022	11,390.28	No	Yes	Yes
304	2022	Oct	10/30/2022	11,452.45	No	Yes	Yes
305	2022	Nov	11/06/2022	10,937.01	No	No	Yes
306	2022	Nov	11/13/2022	11,739.63	No	Yes	Yes
307	2022	Nov	11/20/2022	11,594.64	No	Yes	Yes
308	2022	Nov	11/27/2022	11,838.72	No	Yes	No

As expected, our optimal regression model did worse when supplied the forecasted values from High, Low and Open. It just one actual value within the prediction bounds. But on the other hand, when supplied the actual values for High, Low and Open the regression model did great with a prediction accuracy of 84.62%. We even applied an optimal ARIMA model on the close series since our chosen regression with forecasted values for underlined terms failed. This model gave us a prediction accuracy of 69.23%.

Conclusion:

In summary we conclude that the fitted regression model for Close is still the best model as it has low MSE and includes the effect of High, Low and Open in the model. But at the same time while testing this model failed as we could build a better model for the terms High, Low and Open. We saw that regression did wonders when supplied with actual values for these terms. Hence building better prediction models for High, Low and Open is a future scope.

Problems:

In data collection of the project, we faced issues related to data discrepancy for covid-19 values. Secondly, important variables such as Ticker size and underlying data for the NASDAQ-100 companies was not available without a premium subscription through official NASDAQ website. Another problem faced in model creation for the NASDAQ-100 time series was the non-availability ARIMAX which adds influence of external variables to the response Minitab tool. Although we did try to implement the model in Python, we were successful in understanding and getting the right required results.

Improvements:

The project implementation can potential be improved through the following steps:

- Improvement in models of underlying data for Open, High, and Low
- Addition of weekly company wise NASDAQ stock data to the existing dataset
- More knowledge and Implementation of ARIMAX technique
- Implementation using, Feature-Based Forecast Model Averaging (FFORMA), Recurrent Neural Network (RNN) and LSTM to forecast the NASDAQ 100 index values.

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
High	H
Low	L
Open	O
Total Cases	TC
New Cases	NC
Unemployment rate	UN
Stringency Index	SI

References:

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

https://data.nasdaq.com/tools/python

https://towardsdatascience.com/stock-market-anomalies-and-stock-market-anomaly-detection-are-two-different-things-624331c7b65a

https://finance.yahoo.com/quote/%5ENDX/history/