University of New Brunswick

Time Series Analysis
(EE 6563)

Assignment #2

Professor:
Erik Scheme
Electrical and Computer Engineering

Author:
Saeed Kazemi
(3713280)

March 8, 2021

1. **Explore the attached dataset and find additional information from the resources listed below.**

   (a) The dataset comprises NY Stock Exchange with several additional predictors, as explained in the following paper (especially sections 5 and 6): (Source).

   (b) The original dataset was obtained from: (This link).

   (c) Although the original dataset includes 5 different files, we will only use the "NYSE.csv" file, which includes values from 2010 to 2017

2. **Begin by using your knowledge of ARIMA (or SARIMA) modeling to conduct a univariate time series analysis and prediction of the NY Stock Exchange. Explain your process, present your chosen model, examine the residuals, and evaluate its performance as follows:**

   (a) **Hold out the last 3 months of 2017 for out-of-sample prediction. Plot the predictions and confidence intervals and report the forecasting error using appropriate metrics.**

   (b) **Use a rolling window approach, with a training window of 3 years and daily increments, predict the next day. Again, plot the predictions and confidence intervals and report the forecasting error using appropriate metrics.**

   The general process for Autoregressive Integrated Moving Averages (ARIMA) model is the following:

   (a) *Visualize the Time Series Data (Figure 1)*

   (b) *Transform and make the time series data stationary (Figure 4)*

   (c) *Plot the AutoCorrelation function (ACF) and Partial AutoCorrelation function (PACF) (Figure 3)*

   (d) *Construct the ARIMA Model or Seasonal ARIMA based on the data*

   (e) *Identify optimal model based on Information Criteria, like Akaike Information Critera (AIC)*

   (f) *Evaluate the residuals using statistical test, ACF, PACF, etc. (Figure 5)*

   (g) *Use the model to make predictions (Figure 6)*

   (h) *Calculate forecasting error*

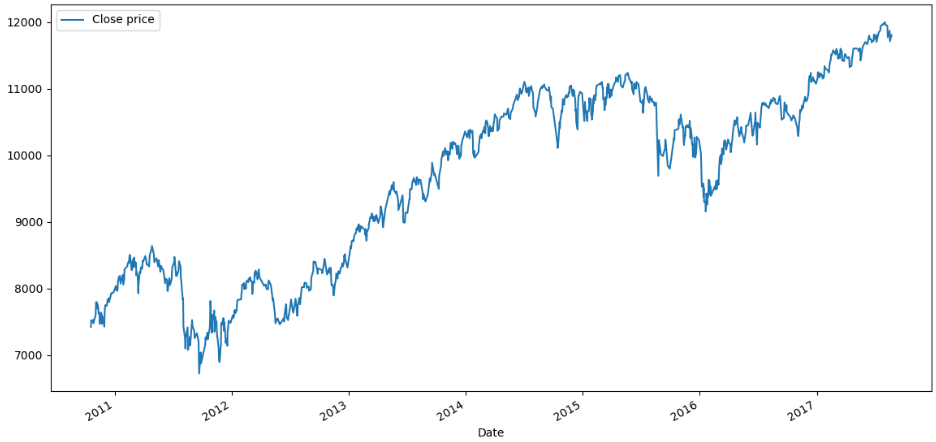

Figure 1: The raw signal of the dataset.

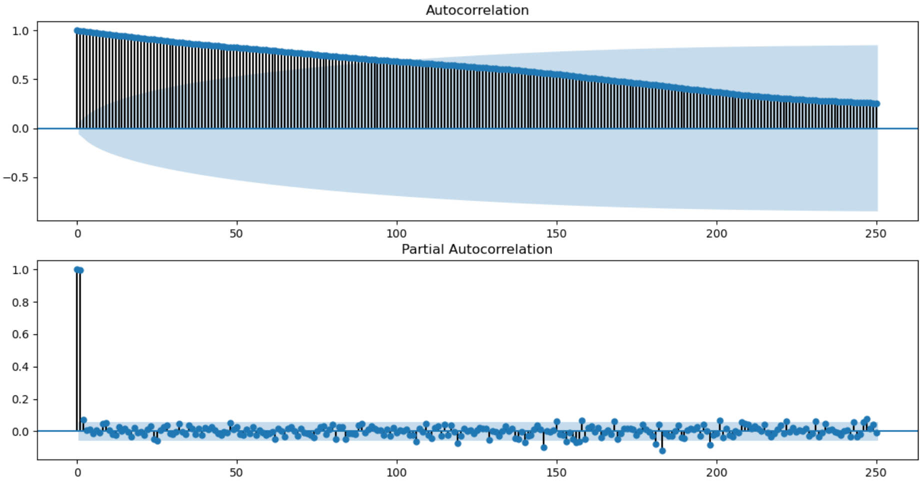

Figure 2: A plot of ACF and PACF of the dataset.

Figures 1 and 2 indicate the raw signal along with its ACF and PACF plots. ACF and PACF plots allow us to determine how correlated points are with each other. Furthermore, the period of the seasonal component can calculate by ACF. As can be seen, the ACF plot has no oscillation. Therefore there is no seasonal component related to our time series.

In order to turn the dataset into a stationary dataset, we used a first-order differencing. Table 1 indicates the result of Augmented Dickey-Fuller test (ADF test) on the first-order differencing. As this test shows, the data got a stationary data. Figures 4 and 3 indicate the 1*st* *order differencing signal along with its ACF and PACF plots. These two plots also show that the data is stationary.*

Table 1: The result of the ADF test on the 1st order differencing in the dataset.

	0
ADF Statistic	-25.690334
p-value	0.000000
#Lags Used	1.000000
Number of Observations Used	1111.000000
Critical Value (1%)	-3.436250
Critical Value (5%)	-2.864145
Critical Value (10%)	-2.568157

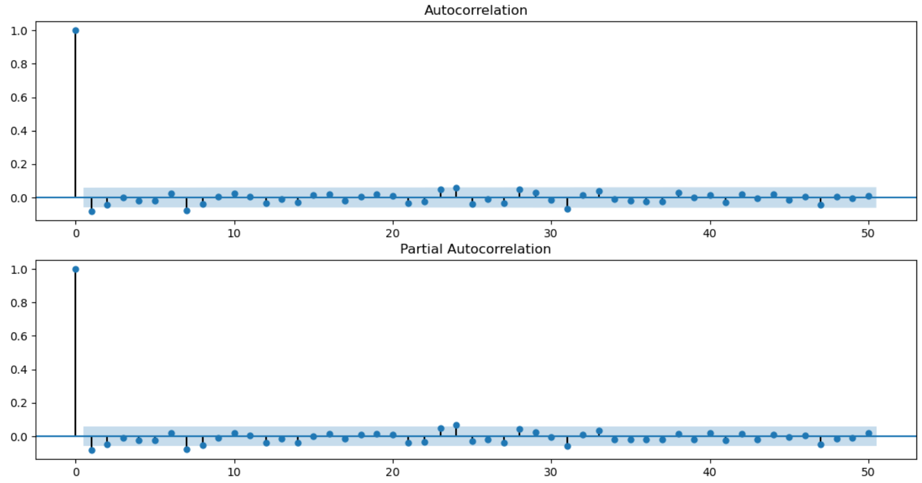

Figure 3: A plot of the ACF and PACF on the 1st order differencing in the dataset.

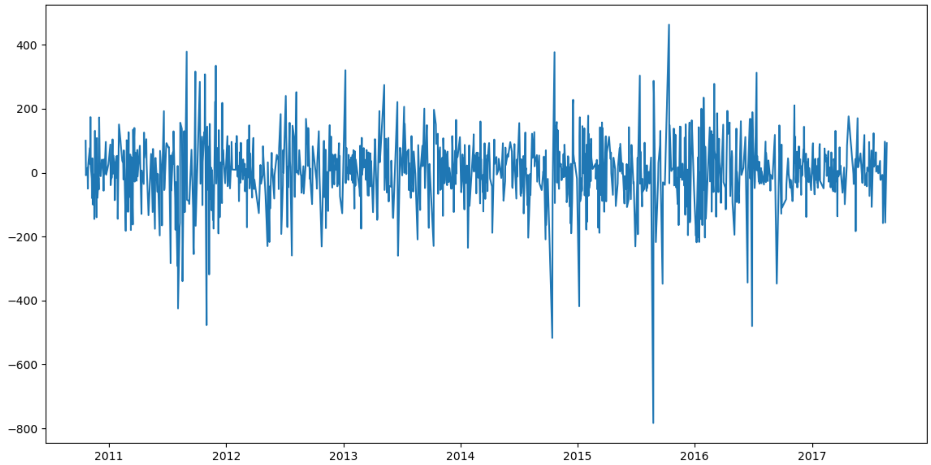

Figure 4: The signal of 1^{st} order differencing.

We used grid search and interpreting ACF and PACF plots to select the best order for ARIMA.

Since both ACF and PACF (see figure 3) had one significant lag, it concluded that both AR order (q) and MA order (p) were one. Also, based on using 1^{st} order differencing, the integrated order (d) set 1.

In grid search, we set a space search to find the model that had the lowest error. Table 2 shows our space search. This method selected the order ARIMA(1, 1, 1) as the best-fitted model.

Table 2: The space search of grid search.

#	hyper-parameter	values
1	p	$1 \sim 4$
2	q	$1 \sim 4$
3	d	1

Figure 5 indicates the residual of the best-fitted model. As this plot illustrates, the residual signal had a Gaussian distribution with zero mean. Besides, the ACF plot shows that this signal was stationary.

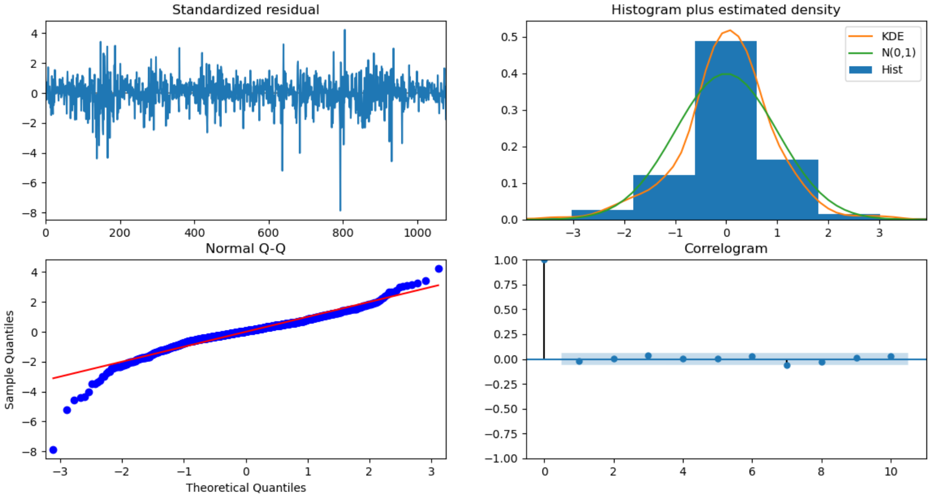

Figure 5: The residual of the best-fitted model (ARIMA(1, 1, 1)) based on grid search.

To predict, we split data into test and train sets. We considered about 35 samples for the test set and others for training. Figure 6 demonstrates the output of the best-fitted model. As can be seen, the model could predict the upward trend. The obtained RMS error was 294.506. Although this model could predict the first sample, it could not follow the rest perfectly.

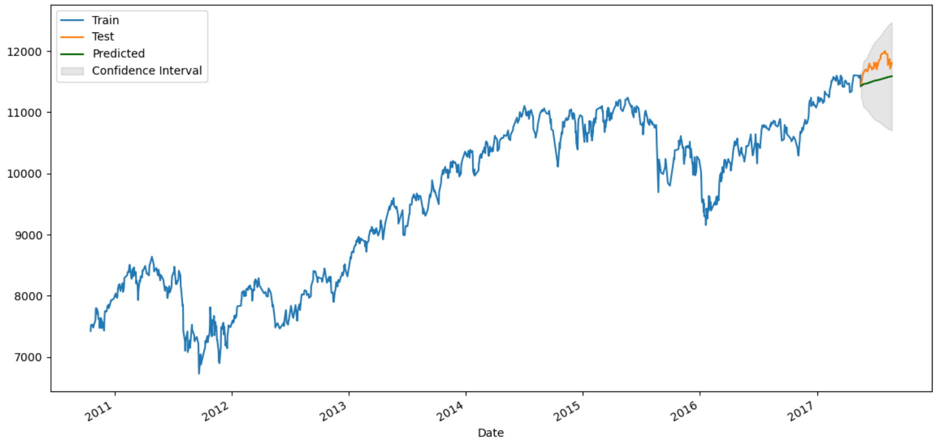

Figure 6: The prediction and actual data of ARIMA(1, 1, 1) model.

For the rolling window approach, we predicted only one day. Then the actual value of this day was added to the training set while our training set had a fixed size. We considered 500 samples (three years). The figure below shows the training set.

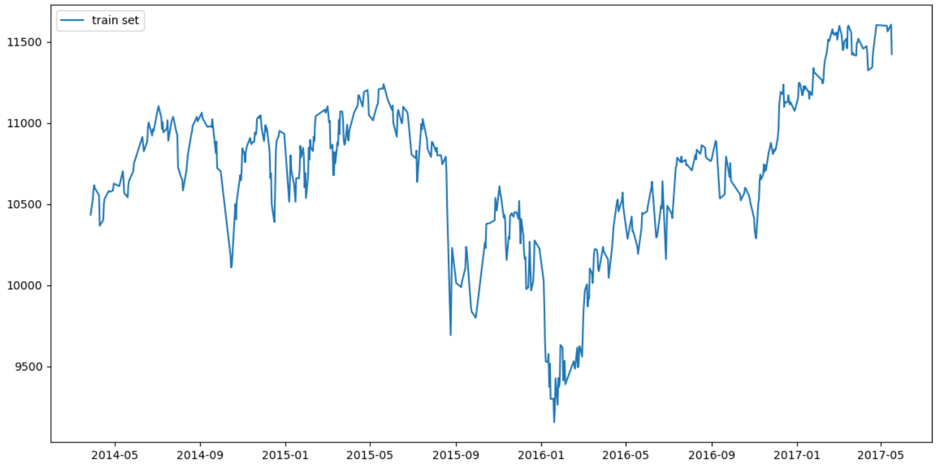

Figure 7: The training set for rolling window approach.

Since the training set did not change, we used the same order of the last model. Figures 9 and 8 demonstrate the output of the rolling window model. As can be seen, the model could follow the test set. Besides, the RMS error reduced from 294.506 to 69.371.

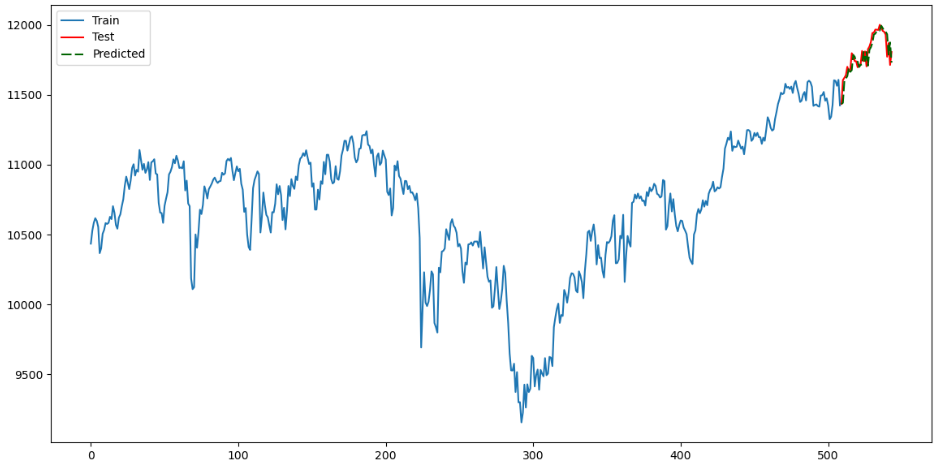

Figure 8: The prediction and actual data of ARIMA$(1, 1, 1)$ model in rolling windows approach.

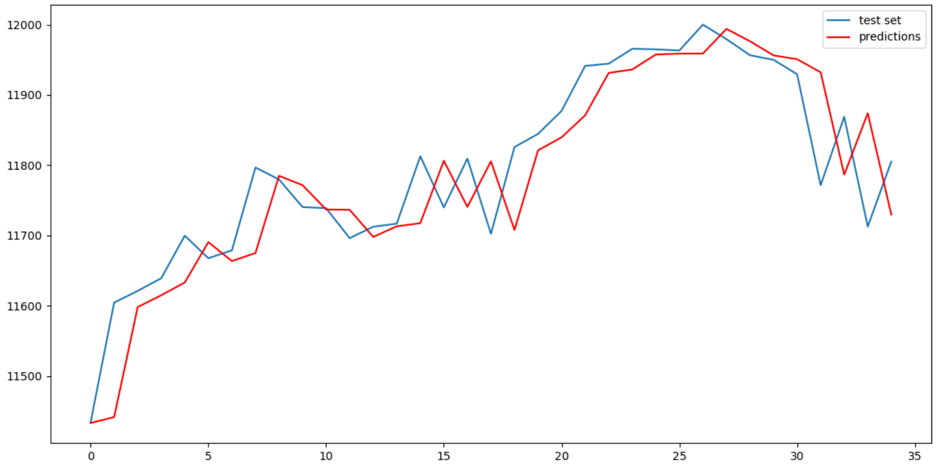

Figure 9: The prediction and test set of ARIMA(1, 1, 1) model in rolling windows approach.

3. **Now, explore whether you can leverage additional information in the file as exogenous variables. Use appropriate tools to evaluate the suitability of using these variables and summarize your results. There is no fixed number of variables assigned – use your judgement and justify your decisions. Using an ARIMAX framework, show the impact of including additional variables as part of your prediction. Repeat the analysis in 2 (a) and (b) using the updated models.**

   First, we visualized the raw signals.

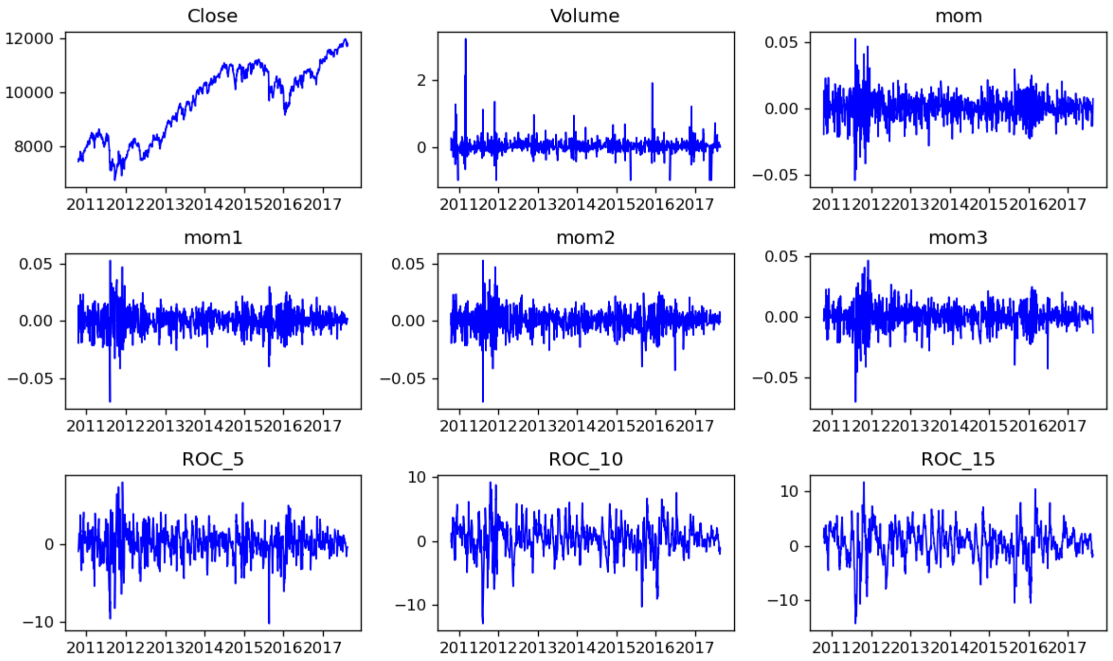

Figure 10: The visualization of the first nine columns.

Before any transformation, we applied the Granger causality test on the first 20 variables. The null hypothesis (H0) of this test says that two variables do not Granger causes, and its alternate hypothesis (H1) indicates the second signal has a significant effect on the first signal. So, wherever P-value is less than 0.05, we can consider Grange causality. Table 3 indicates the result of this test on the first twelve columns. For this test, the maximum lag set 12.
Based on this test, signals mom, mom1, ROC_X, Oil, and EMA_X ~~were Granger causality of~~ the "Close price"; We stored these variables for the next steps.

Table 3: The result of Granger causality test.

#	Variable Name	min_p_value	lag	Causality
1	Close	1.0000	1.0	-
2	Volume	0.2613	10.0	-
3	mom	0.0022	8.0	✓
4	mom1	0.0285	7.0	✓
5	mom2	0.2049	2.0	-
6	mom3	0.0860	1.0	-
7	ROC_5	0.0024	1.0	✓
8	ROC_10	0.0036	1.0	✓
9	ROC_15	0.0037	1.0	✓
10	ROC_20	0.0010	9.0	✓
11	EMA_10	0.0021	1.0	✓
12	EMA_20	0.0014	1.0	✓
13	EMA_50	0.0008	9.0	✓
14	EMA_200	0.0035	9.0	✓
15	DTB4WK	0.2096	1.0	-
16	DTB3	0.1918	1.0	-
17	DTB6	0.2826	10.0	-
18	DGS5	0.5876	5.0	-
19	DGS10	0.3805	5.0	-
20	Oil	0.0120	7.0	✓

As these signals had a different range, we used standardization to transform all data to scale -1 and 1. Figure 11 indicates the standardized variables. Also, we applied ADF test on these signals to find out that these signals were stationary or non-stationary. Table 4 shows the result of this test on the dataset.

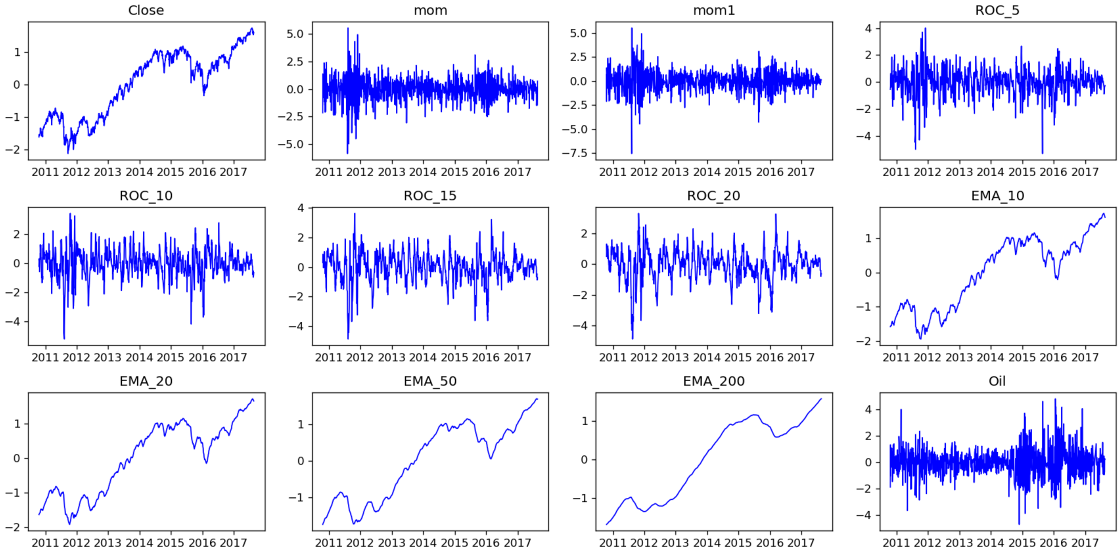

Figure 11: The visualization of the first twelve columns. Columns were standardized.

Table 4: The result of the ADF test on the dataset.

	Close	mom	mom1	ROC_5	ROC_10	ROC_15	ROC_20	EMA_10	EMA_20	EMA_50	EMA_200	Oil
ADF Statistic	-1.085	-12.134	-38.465	-10.619	-9.753	-6.995	-8.074	-0.967	-0.882	-0.791	-0.585	-8.920
p-value	0.7	1.7e-22	0	5.5e-19	7.9e-17	7.5e-10	1.5e-12	0.7	0.7	0.8	0.8	1.0e-14
# Lags Used	2	11	0	9	6	16	5	3	3	6	7	10
# Obs. Used	1111	1102	1113	1104	1107	1097	1108	1110	1110	1107	1106	1103
Critical Value (1%)	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436
Critical Value (5%)	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864
Critical Value (10%)	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568

According to table 4 all columns were stationary except the "Close price" and EMA_X variables. Therefore, we used a first-order differencing to turn these time series into the stationary data. Figures 12 and 13 indicate this 1^{st} order differencing on "Close price" along with its ACF and PACF plots. These two plots also show that the data got stationary. Also, table 5 shows the ADF test on the 1^{st} order differencing variables.

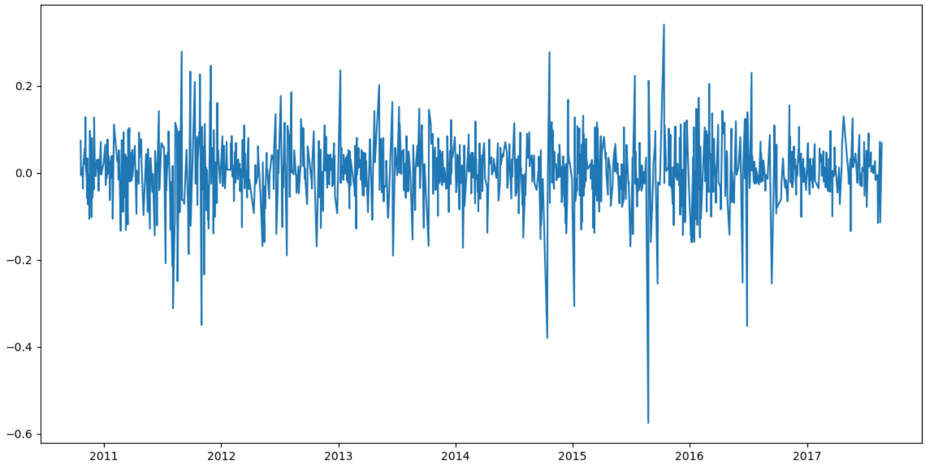

Figure 12: The 1^{st} order differencing of ”Close price” data.

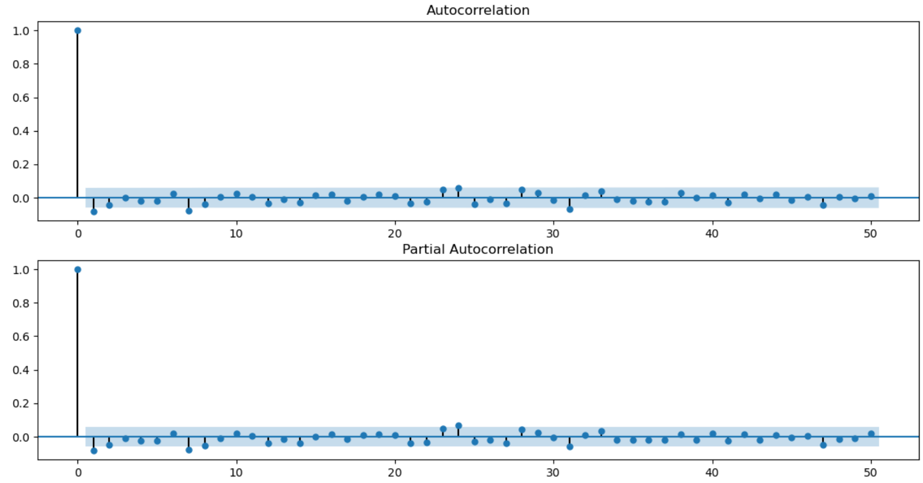

Figure 13: A plot of ACF and PACF on the 1^{st} order differencing in ”Close price”.

Table 5: The result of the ADF test on the 1^{st} order differencing variables.

	Close	mom	mom1	ROC_5	ROC_10	ROC_15	ROC_20	EMA_10	EMA_20	EMA_50	EMA_200	Oil
ADF Statistic	-25.69	-12.13	-24.15	-10.63	-9.74	-6.99	-8.059	-13.44	-10.72	-6.82	-4.72	-8.91
p-value	0	1.7e-22	0	4.9e-19	8.1e-17	7.4e-10	1.6e-12	3.7e-25	3.0e-19	1.9e-09	7.6e-5	1.0e-14
#Lags Used	1	11	1	9	6	16	5	2	2	5	6	10
# Observations Used	1111	1101	1111	1103	1106	1096	1107	1110	1110	1.107	1106	1102
Critical Value (1%)	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436
Critical Value (5%)	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864
Critical Value (10%)	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568

After transforming and making stationary, we needed to shift exogenous variables as much as the corresponding lag in the Granger causality result.

We fitted our model based on exogenous variables. The order of the model for this question was similar to the last question. Figure 14 indicates the residual of the fitted model. As this plot illustrates, the residual signal of the model had a Gaussian distribution with zero mean. Besides, the ACF plot shows that this signal is stationary.

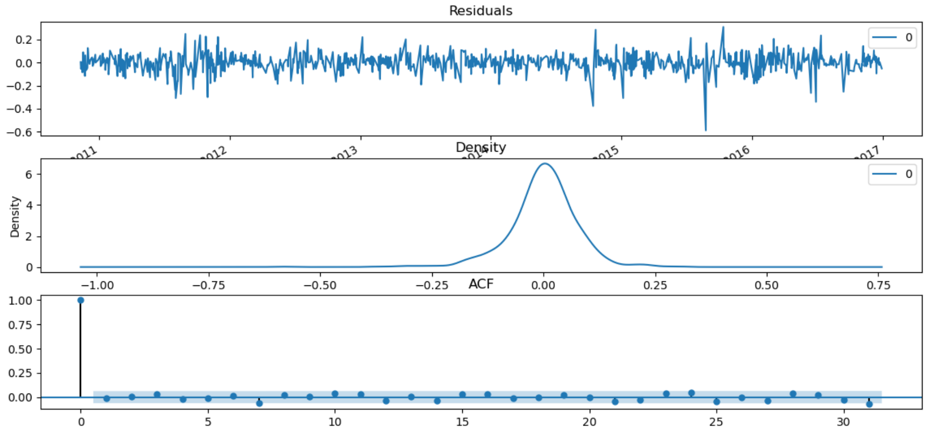

Figure 14: The residual of the fitted model (ARIMAX(1, 0, 1)).

To forecast, we split data into test and train set like part 2. After prediction, we needed to reverse all transformations to get the real forecast. Figure 15 demonstrates the output of the fitted model. As can be seen, because of using the exogenous variables, the model could predict some changes, and for others could not follow the actual data. The RMS error was 328.647.

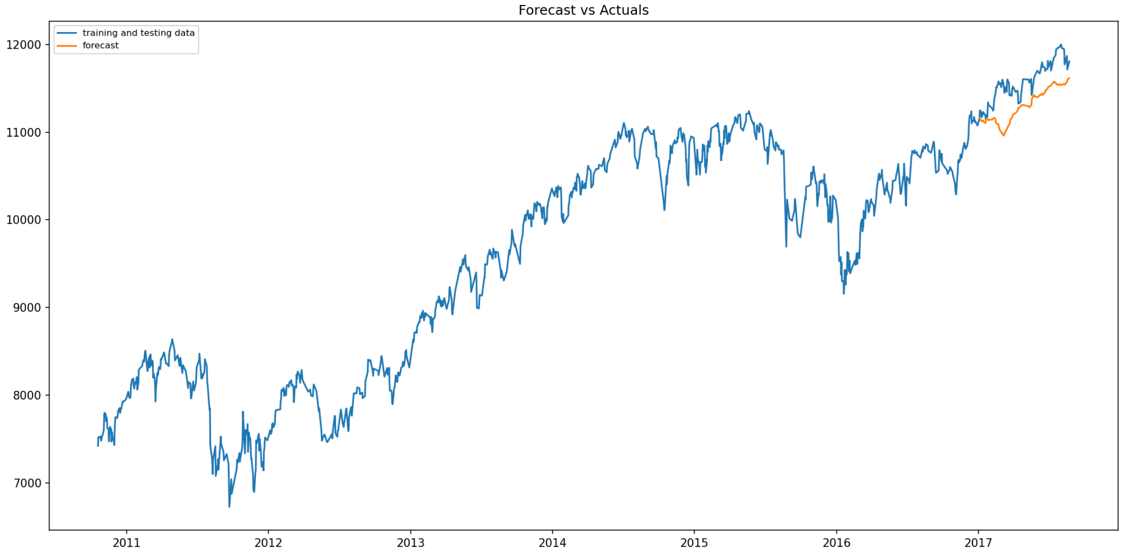

Figure 15: The prediction and actual data of ARIMAX(1, 0, 1) model.

For the rolling window approach, we used the same procedure in question 2. Figure 16 demonstrates the output of the rolling window model for ARIMAX. As can be seen, the model could follow the test set better than previous model. The RMS error decreased from 328.647 to 256.297. It seems that the prediction of both models had a negative bias.

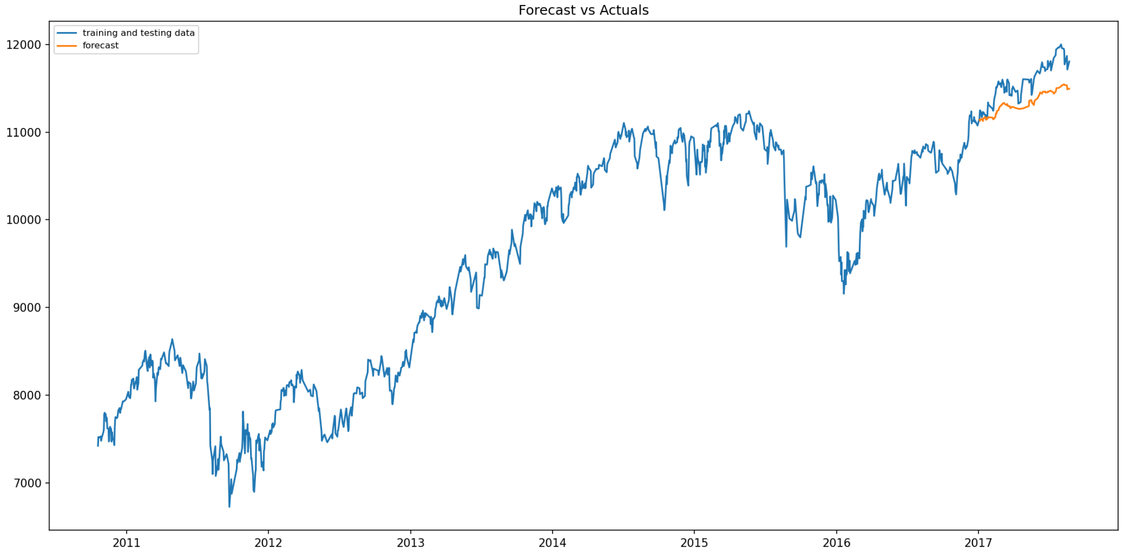

Figure 16: The prediction and actual data of ARIMAX(1, 0, 1) model in rolling windows approach.

4. **Finally, repeat the analysis in 3 but for a VAR-based approach. Explore the relationships between the chosen variables using appropriate tools. Again, there is no fixed number of variables assigned – use your judgement and justify your decisions. Using a VAR framework, show the impact of including additional variables in your prediction. Repeat the analysis in 2 (a) and (b) using the updated models.**

   Likewise the last question, we visualized the first nine signals to see what kind of data we have.

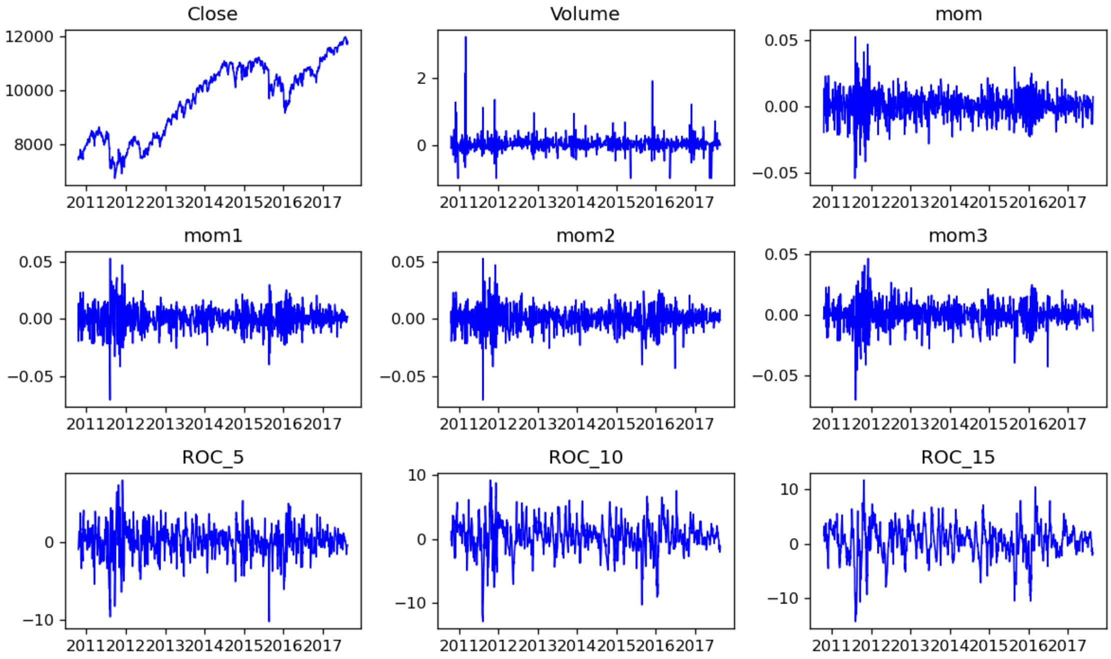

Figure 17: The visualization of the first nine columns.

Then, we needed to find out which variables must choose for VAR, so that we applied two tools, Granger causality and Johanson's Cointegration test. Granger causality test implemented over the first 20 variables. Then we used the second test on variables that were Granger causes. Tables 6 and 7 show the results of these two tests. Based on these tests, signals mom, mom1, ROC_X and EMA_X had a relation with the target variable ("Close price"). We stored these variables for the next steps.

Table 6: The result of Granger causality test.

#	Variable Name	min_p_value	lag	Causality
1	Close	1.0000	1.0	-
2	Volume	0.2613	10.0	-
3	mom	0.0022	8.0	✓
4	mom1	0.0285	7.0	✓
5	mom2	0.2049	2.0	-
6	mom3	0.0860	1.0	-
7	ROC_5	0.0024	1.0	✓
8	ROC_10	0.0036	1.0	✓
9	ROC_15	0.0037	1.0	✓
10	ROC_20	0.0010	9.0	✓
11	EMA_10	0.0021	1.0	✓
12	EMA_20	0.0014	1.0	✓
13	EMA_50	0.0008	9.0	✓
14	EMA_200	0.0035	9.0	✓
15	DTB4WK	0.2096	1.0	-
16	DTB3	0.1918	1.0	-
17	DTB6	0.2826	10.0	-
18	DGS5	0.5876	5.0	-
19	DGS10	0.3805	5.0	-
20	Oil	0.0120	7.0	✓

Table 7: The result of Johanson's Cointegration test.

#	Variable Name	Test Stat	C(95%)	Signif
1	Close	2460.9	311.1280	True
2	mom	1996.6	263.2603	True
3	mom1	1666.7	219.4051	True
4	ROC_5	1340.7	179.5190	True
5	ROC_10	1031.2	143.6690	True
6	ROC_15	804.2	111.7797	True
7	ROC_20	628.4	83.9383	True
8	EMA_10	459.3	60.0627	True
9	EMA_20	305.6	40.1749	True
10	EMA_50	183.3	24.2761	True
11	EMA_200	81.8	12.3212	True
12	Oil	2.7	4.1296	False

As figure 17 shown, signals had a different range. Therefore, we standardized all data to have the same scale. Figure 18 indicates the standardized variables. Also, we applied ADF test on these signals to find out which were stationary or non-stationary. Table 8 shows the result of this test on the dataset.

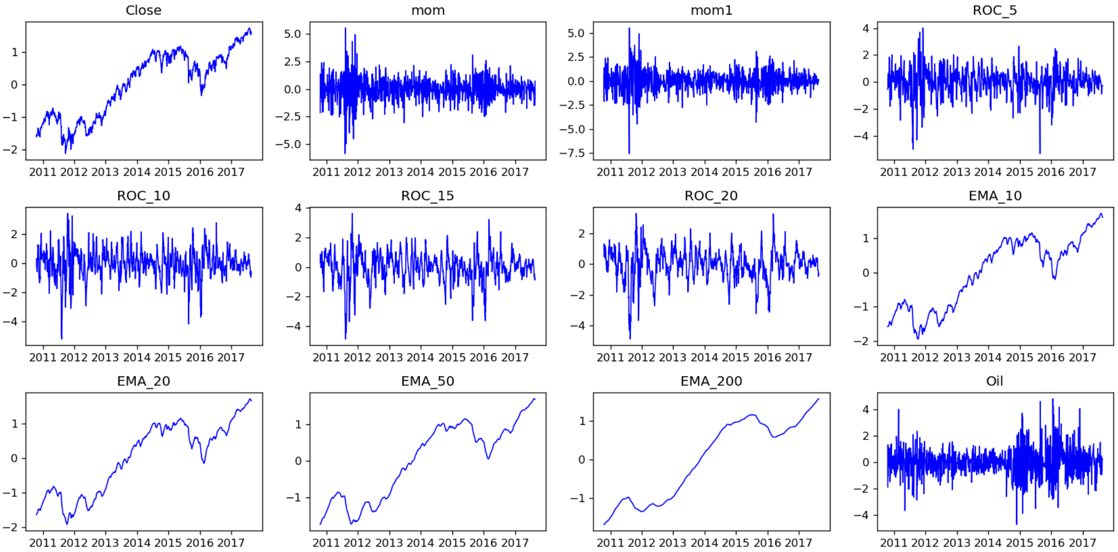

Figure 18: The visualization of the first twelve columns. Columns were standardized.

Table 8: The result of the ADF test on the dataset.

	Close	mom	mom1	ROC_5	ROC_10	ROC_15	ROC_20	EMA_10	EMA_20	EMA_50	EMA_200	Oil
ADF Statistic	-1.085	-12.134	-38.465	-10.619	-9.753	-6.995	-8.074	-0.967	-0.882	-0.791	-0.585	-8.920
p-value	0.7	1.7e-22	0	5.5e-19	7.9e-17	7.5e-10	1.5e-12	0.7	0.7	0.8	0.8	1.0e-14
# Lags Used	2	11	0	9	6	16	5	3	3	6	7	10
# Obs. Used	1111	1102	1113	1104	1107	1097	1108	1110	1110	1107	1106	1103
Critical Value (1%)	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436
Critical Value (5%)	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864
Critical Value (10%)	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568

According to table 8, all columns were stationary except "Close price" and EMA_X variables. We used a first-order differencing to turn these time series into the stationary data. Figures 19 and 20 indicate this 1^{st} *order differencing for the "Close price" signal along with its ACF and PACF plots. These two plots also show that the data got stationary. Table 9 shows the ADF test on the* 1^{st} *order differencing variables.*

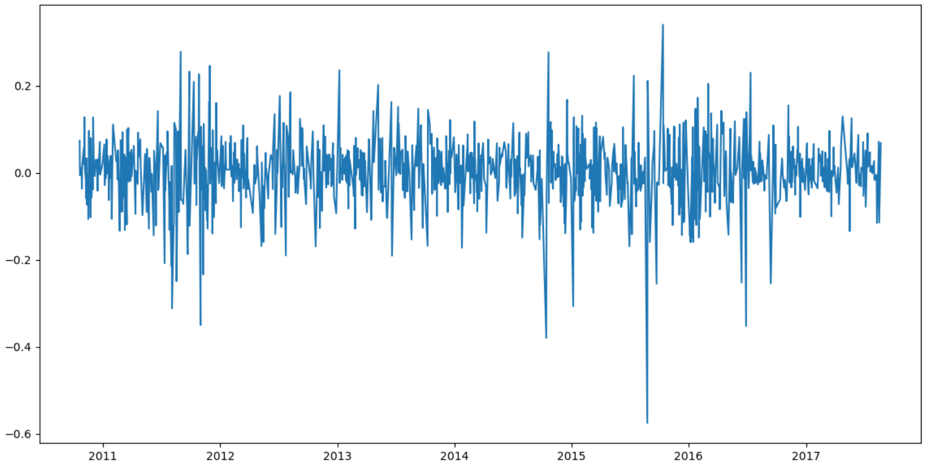

Figure 19: The 1^{st} order differencing of ”Close price” data.

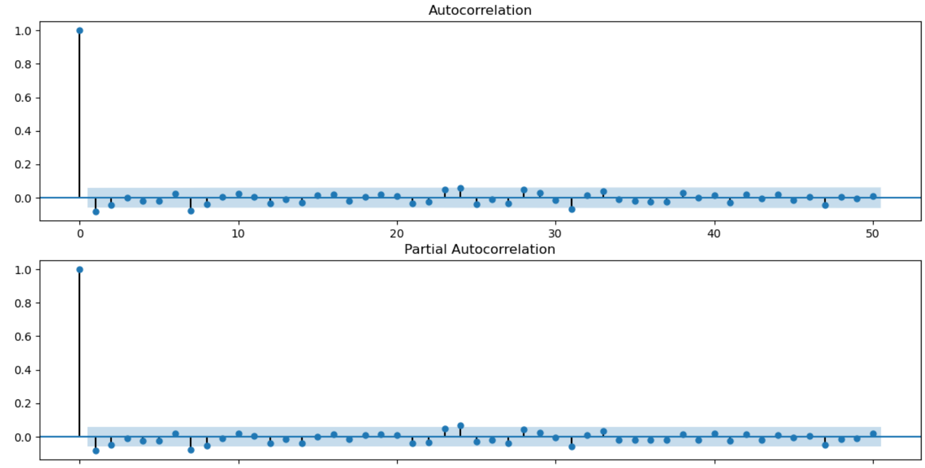

Figure 20: A plot of ACF and PACF on the 1^{st} order differencing in ”Close price”.

Table 9: The result of the ADF test on the 1^{st} order differencing variables.

	Close	mom	mom1	ROC_5	ROC_10	ROC_15	ROC_20	EMA_10	EMA_20	EMA_50	EMA_200	Oil
ADF Statistic	-25.69	-12.13	-24.15	-10.63	-9.74	-6.99	-8.059	-13.44	-10.72	-6.82	-4.72	-8.91
p-value	0	1.7e-22	0	4.9e-19	8.1e-17	7.4e-10	1.6e-12	3.7e-25	3.0e-19	1.9e-09	7.6e-5	1.0e-14
#Lags Used	1	11	1	9	6	16	5	2	2	5	6	10
# Observations Used	1111	1101	1111	1103	1106	1096	1107	1110	1110	1.107	1106	1102
Critical Value (1%)	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436	-3.436
Critical Value (5%)	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864	-2.864
Critical Value (10%)	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568	-2.568

After transforming and making stationary, we needed to find the best-fitted model. Thus we searched on various orders. For the searching area, we used all lags between 1 and 50. Figure 21 shows the AIC over the different orders. The best order in terms of AIC belonged to order 9 (VAR(9) model). We could do the same procedure to find the best order based on other information criteria. The below table shows the result of other information criteria.

Table 10: The result of other information criteria.

#	Information Criteria	Best order
1	AIC	9
2	BIC	3
3	FPE	9
4	HQIC	5

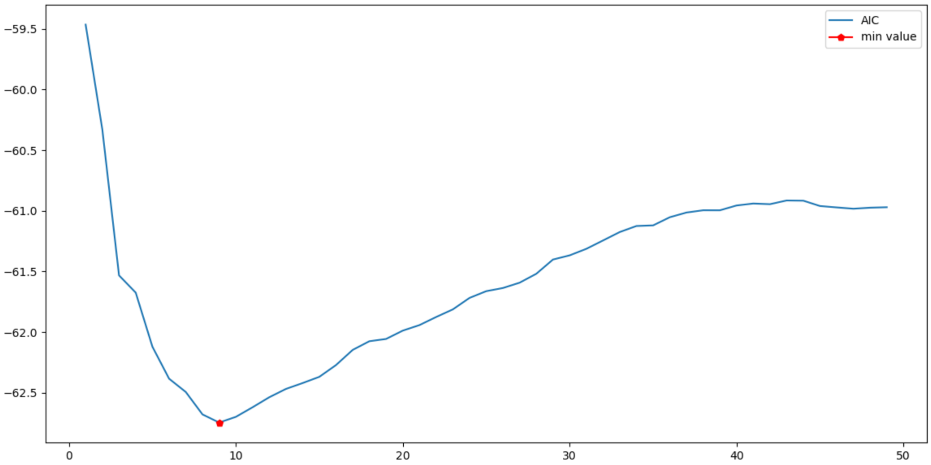

Figure 21: The AIC over the different orders of VAR model.

Figure 22 indicates the residual of the fitted model only for the first variable. As this plot illustrates, the residual signal of the model had a Gaussian distribution with zero mean. Besides, the ACF plot shows that this signal was stationary.

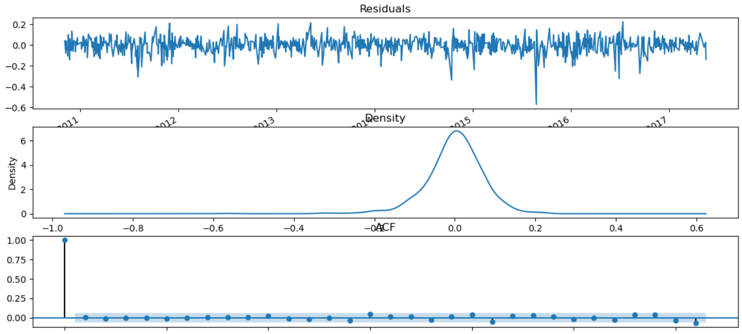

Figure 22: The residual of the fitted model.

In addition, we applied the Durbin Watson test to estimate the correlation between residuals. This correlation shows whether or not the pattern has any leftover in the residuals. As table 11 indicates Durbin Watson's results were close to 2. It means there was no correlation between residuals.

Table 11: The result of Durbin Watson test on the dataset.

#	Variable Name	The result
1	Close	1.99
2	mom	1.98
3	mom1	2.0
4	ROC_5	1.99
5	ROC_10	1.99
6	ROC_15	1.97
7	ROC_20	1.97
8	EMA_10	2.0
9	EMA_20	2.0
10	EMA_50	2.0
11	EMA_200	2.0
12	Oil	2.0

To forecast, we split data into test and train set like the last two parts. After prediction, we reversed differencing and standardization to get the real forecast. Figure 23 demonstrates the output of the fitted model only for "Close price." As can be seen, although the model could predict perfectly at the first, it got a flat slope. The RMS error for this model was 339.599.

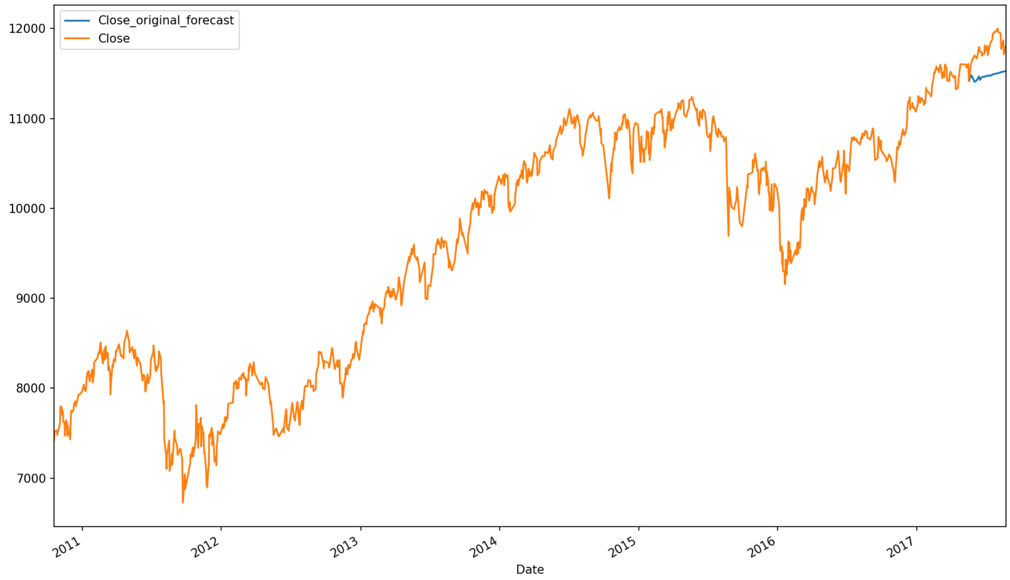

Figure 23: The prediction and actual data of VAR model.

For the rolling window approach, we used the same procedure in question 2. Figure 24 demonstrates the output of the rolling window model for VAR. As can be seen, the model could follow the test set better than the previous models. The RMS error decreased from 339.599 to 117.565.

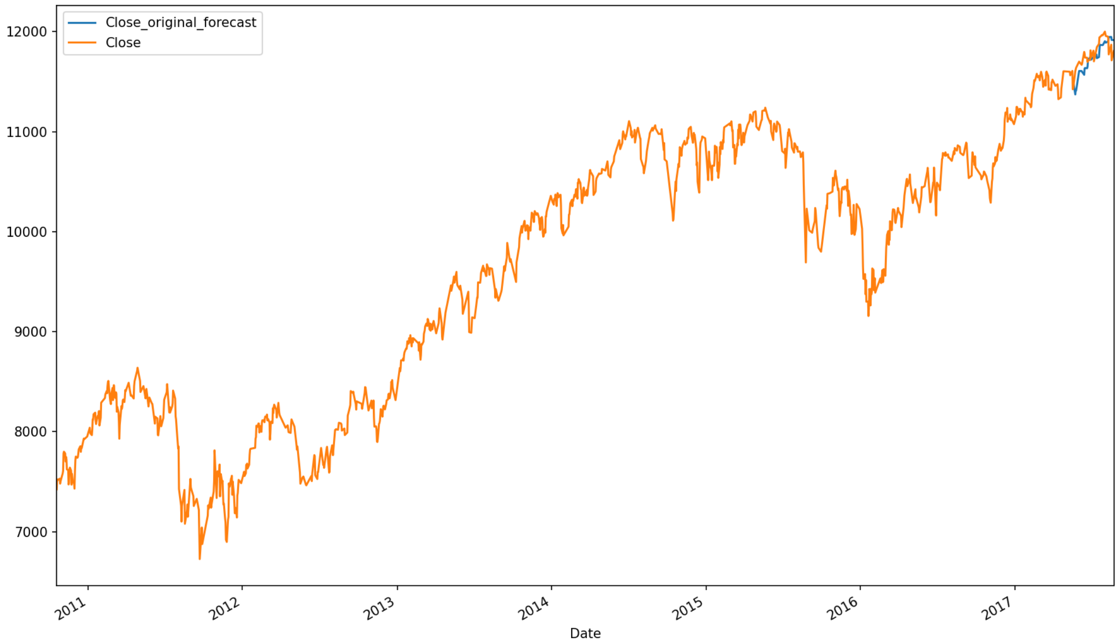

Figure 24: The prediction and actual data of VAR model.