4/18/2023

Using ARIMA to Forecast Copper Future Prices

Time Series Analysis

Contents

About the Data

The prices of futures contract on copper traded on MCX (Multi Commodity Exchange) from December 2011 to March 2021 have been used as the dataset in this report. The prices are expressed in INR. The data is sourced from Investing.com.

Design of the Study

We have 2491 observations in total. We split the entire dataset into training (first 2000) observations and testing (remaining 491) observations. We start with the first 2000 observations and predict the 2001th observation on that basis. To predict the 2002th observation, we take the next 2000 observations (2nd to 2001th observation). This way each observation in the testing set is predicted using the previous 2000 observations on a rolling basis. The initial exploratory data analysis (EDA) is done on the first 2000 observations only.

Test for Stationarity

A data generating process is said to be stationary when its statistical properties do not change with time. In other words, a stationary series is one whose mean, variance or autocorrelation structure does not vary with time. Graphically, it has a randomly fluctuating and horizontal plot (refer to figure 2). In order to check for stationarity, the Augmented Dickey-Fuller (ADF) test is used. Here, we run an ADF Test on the log of copper future prices in the training set to check if the series is stationary. The series is found to be non-stationary in nature. We then take a first difference of the series and recheck for stationarity. The first difference of the log of copper prices is found to be stationary in nature. Since we have arrived at a stationary series by taking a difference once, the series is said to be integrated of order 1 or I(1). Below is a visual representation of the series and its first difference.

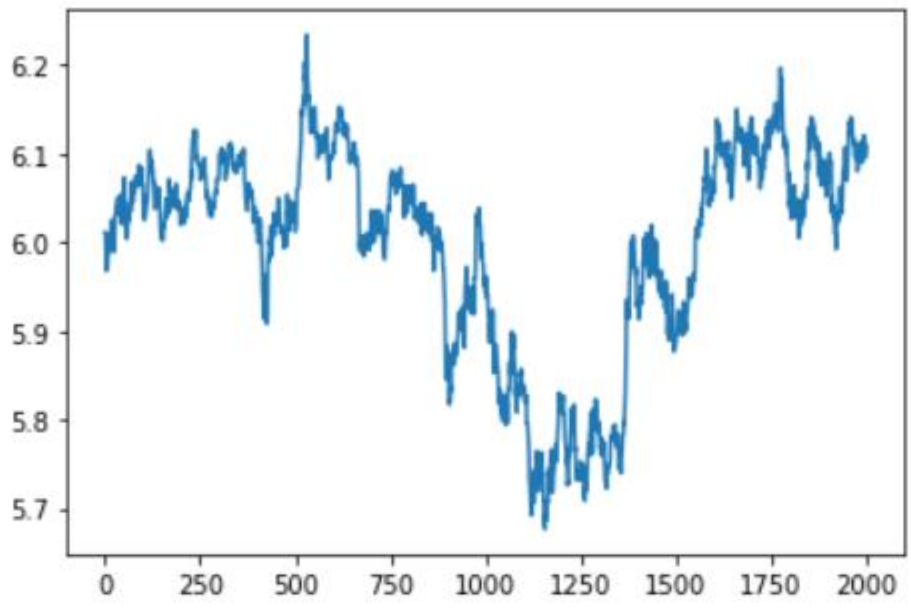


Figure 1: Visual Representation of Natural Log of Copper Futures

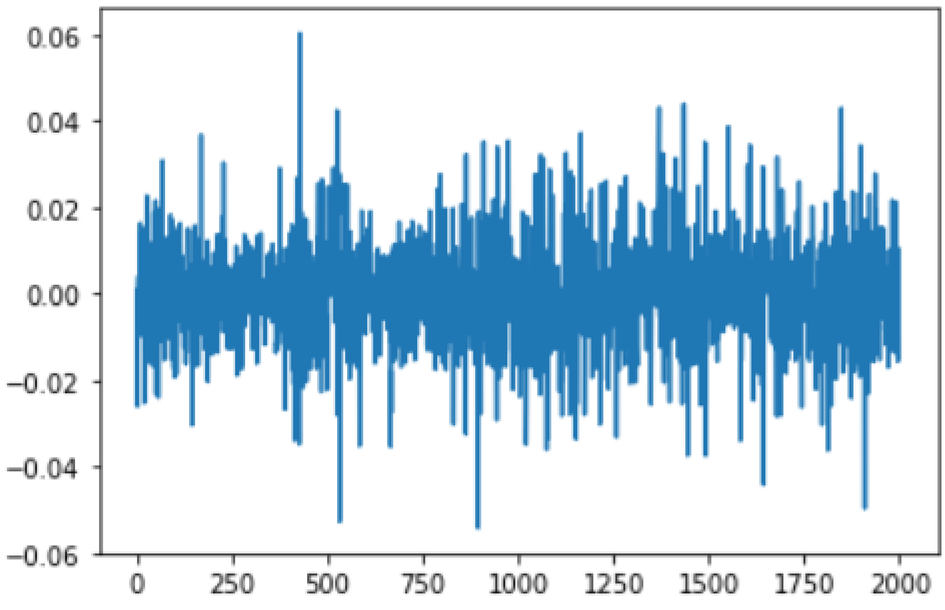


Figure 2: Visual Representation of First Difference of Series in Figure 3

ACF and PACF on First Difference

A correlogram is plotted to identify the appropriate lags till which a significant autocorrelation is observed in the stationary first difference. The appropriate lag in both the plots will help in determination of the appropriate ARIMA Model. In case of both the ACF and PACF, a significant negative correlation is found at lag 16. This implies that 16 is the appropriate value of both ‘p’ and ‘q’, which form 2 of the parameters of the ARIMA(p, d, q) model. It should be notes that the correlation values in both plots are significantly different from 0, but very weak by themselves.

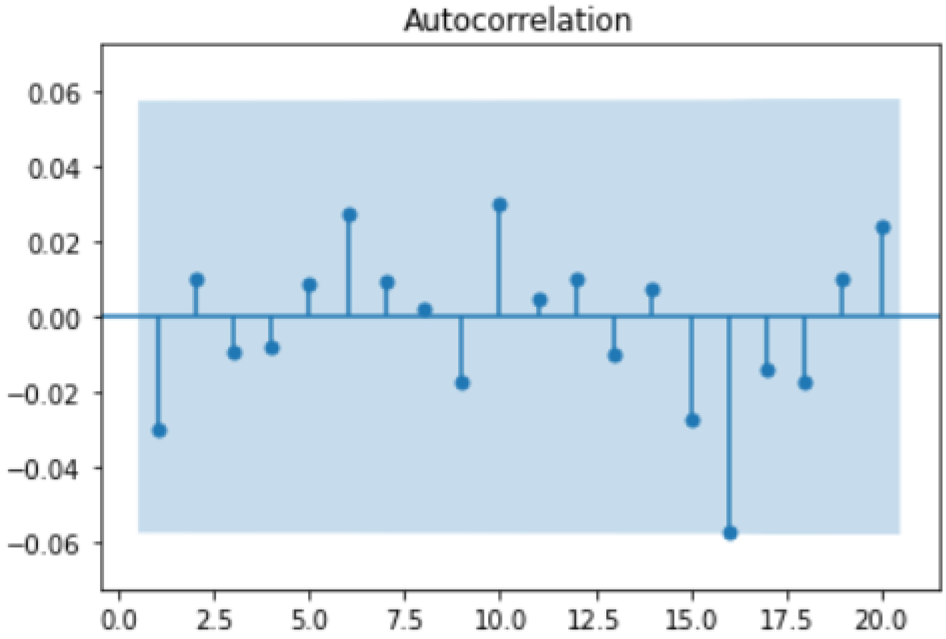


Figure 3: ACF of First Difference

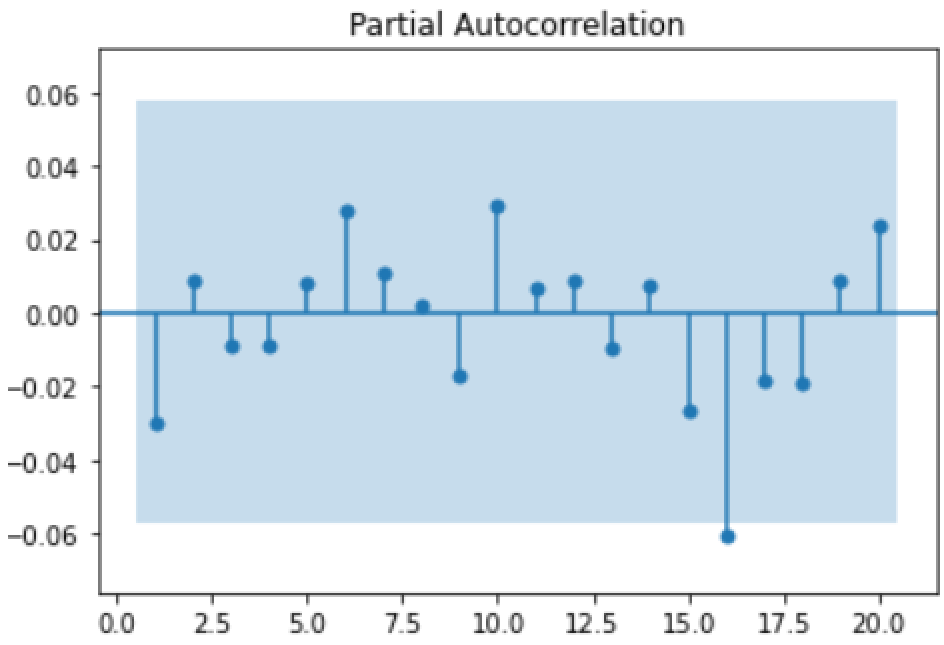


Figure 4: PACF of First Difference

Research Objective

The objective of this research is to evaluate the prediction capability of ARIMA models on financial time series (prices of commodity futures). In order to achieve this object, we evaluate the performance of ARIMA models in the testing dataset. We also try to discover the optimal parameters for the ARIMA model using both statistical and machine learning based approaches.

ARIMA – statistically determined hyper-parameters

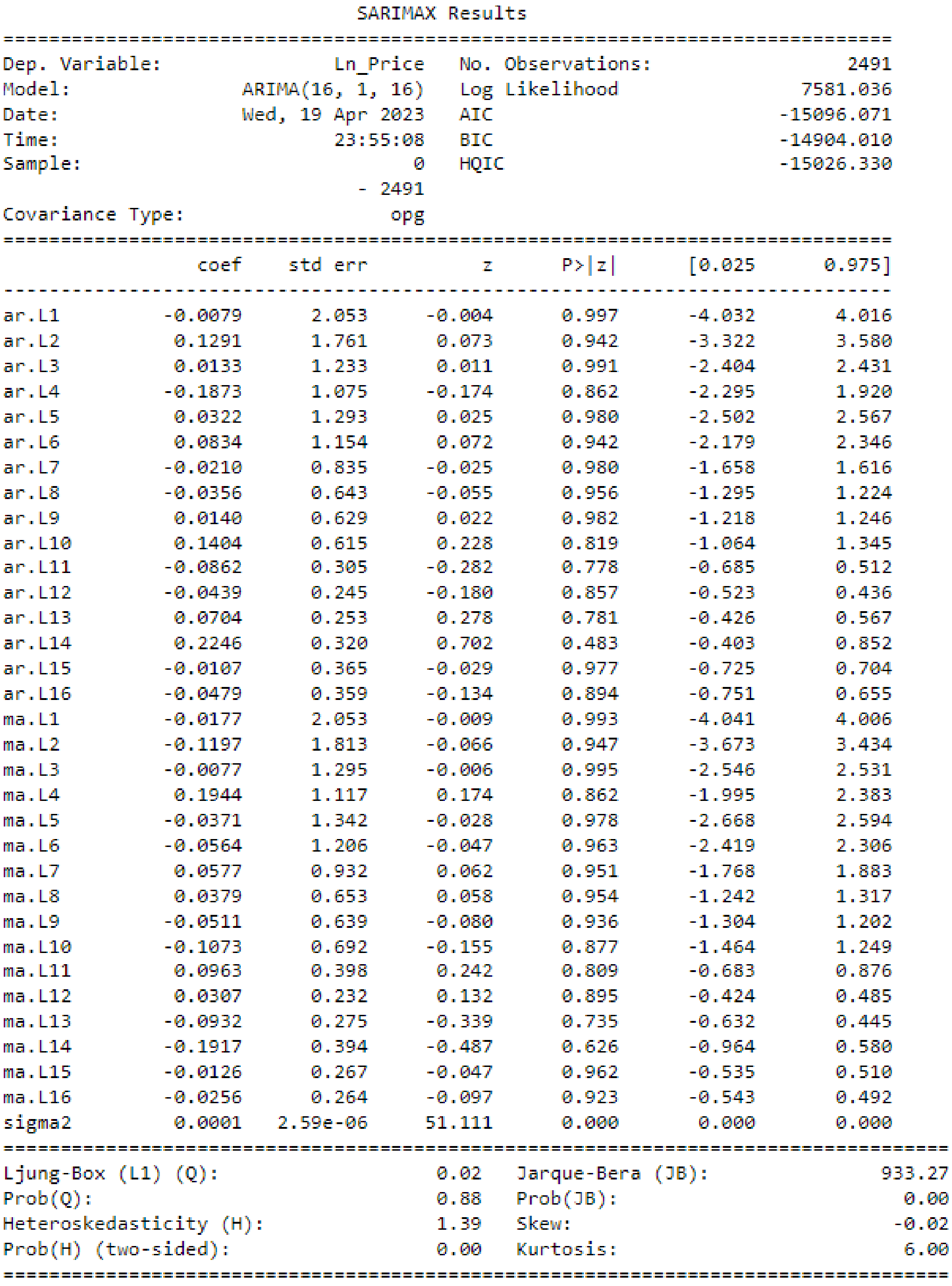
ARIMA models are a combination of Autoregressive (AR) models and Moving Average (MA) models. AR models are fit by regressing a time series variable with its lagged self. MA models are fit by regressing a time series variable with the lagged errors of that model. The lag parameter for the AR model is denoted by ‘p’ and that of the MA model is denoted by ‘q’. The ‘I’ in ARIMA stands for ‘Integrated’ which refers to the degree of differentiation of the target variable to make it stationary. After conducting the above analysis, we arrive at the appropriate order of differencing

(1), the appropriate lag value for the AR component of ARIMA (16) and the appropriate lag term for the MA component of ARIMA (16). We therefore use the ARIMA(16, 1, 16) model. The comprehensive model equation is as follows.

$$Y_t = c + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_{16} Y_{t-16} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_{16} \varepsilon_{t-16} + \varepsilon_t$$

Where Y_t represents the first difference of natural logarithm of Copper Future Prices. The fit of the above model is described below.

Table 1: Estimates Table



```
                               SARIMAX Results
==============================================================================
Dep. Variable:               Ln_Price   No. Observations:                 2491
Model:               ARIMA(16, 1, 16)   Log Likelihood                7581.036
Date:                Wed, 19 Apr 2023   AIC                         -15096.071
Time:                        23:55:08   BIC                         -14904.010
Sample:                             0   HQIC                        -15026.330
                               - 2491
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0079      2.053     -0.004      0.997      -4.032       4.016
ar.L2          0.1291      1.761      0.073      0.942      -3.322       3.580
ar.L3          0.0133      1.233      0.011      0.991      -2.404       2.431
ar.L4         -0.1873      1.075     -0.174      0.862      -2.295       1.920
ar.L5          0.0322      1.293      0.025      0.980      -2.502       2.567
ar.L6          0.0834      1.154      0.072      0.942      -2.179       2.346
ar.L7         -0.0210      0.835     -0.025      0.980      -1.658       1.616
ar.L8         -0.0356      0.643     -0.055      0.956      -1.295       1.224
ar.L9          0.0140      0.629      0.022      0.982      -1.218       1.246
ar.L10         0.1404      0.615      0.228      0.819      -1.064       1.345
ar.L11        -0.0862      0.305     -0.282      0.778      -0.685       0.512
ar.L12        -0.0439      0.245     -0.180      0.857      -0.523       0.436
ar.L13         0.0704      0.253      0.278      0.781      -0.426       0.567
ar.L14         0.2246      0.320      0.702      0.483      -0.403       0.852
ar.L15        -0.0107      0.365     -0.029      0.977      -0.725       0.704
ar.L16        -0.0479      0.359     -0.134      0.894      -0.751       0.655
ma.L1         -0.0177      2.053     -0.009      0.993      -4.041       4.006
ma.L2         -0.1197      1.813     -0.066      0.947      -3.673       3.434
ma.L3         -0.0077      1.295     -0.006      0.995      -2.546       2.531
ma.L4          0.1944      1.117      0.174      0.862      -1.995       2.383
ma.L5         -0.0371      1.342     -0.028      0.978      -2.668       2.594
ma.L6         -0.0564      1.206     -0.047      0.963      -2.419       2.306
ma.L7          0.0577      0.932      0.062      0.951      -1.768       1.883
ma.L8          0.0379      0.653      0.058      0.954      -1.242       1.317
ma.L9         -0.0511      0.639     -0.080      0.936      -1.304       1.202
ma.L10        -0.1073      0.692     -0.155      0.877      -1.464       1.249
ma.L11         0.0963      0.398      0.242      0.809      -0.683       0.876
ma.L12         0.0307      0.232      0.132      0.895      -0.424       0.485
ma.L13        -0.0932      0.275     -0.339      0.735      -0.632       0.445
ma.L14        -0.1917      0.394     -0.487      0.626      -0.964       0.580
ma.L15        -0.0126      0.267     -0.047      0.962      -0.535       0.510
ma.L16        -0.0256      0.264     -0.097      0.923      -0.543       0.492
sigma2         0.0001   2.59e-06     51.111      0.000       0.000       0.000
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               933.27
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               1.39   Skew:                            -0.02
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.00
===================================================================================
```

ARIMA – Machine Learning based tuning

As it was observed from the ACF and PACF plots, while significant (non 0) autocorrelation was detected at lag 16, the correlation itself was not strong enough. Therefore, we adopt an alternative approach to determining the hyper parameters of the ARIMA(p,d,q) model. It should be noted that 'd' is already known to be 1, as the time series variable is Integrated of order 1. In order to select the hyper-parameters, we fit project the values in the testing period, compute the root mean squared error (RMSE) between the projected and the actual observations in the test set and select the model which has least RMSE. We vary the values of ‘p’ and ‘q’ from 1 to 4 (total 16 possible combinations).

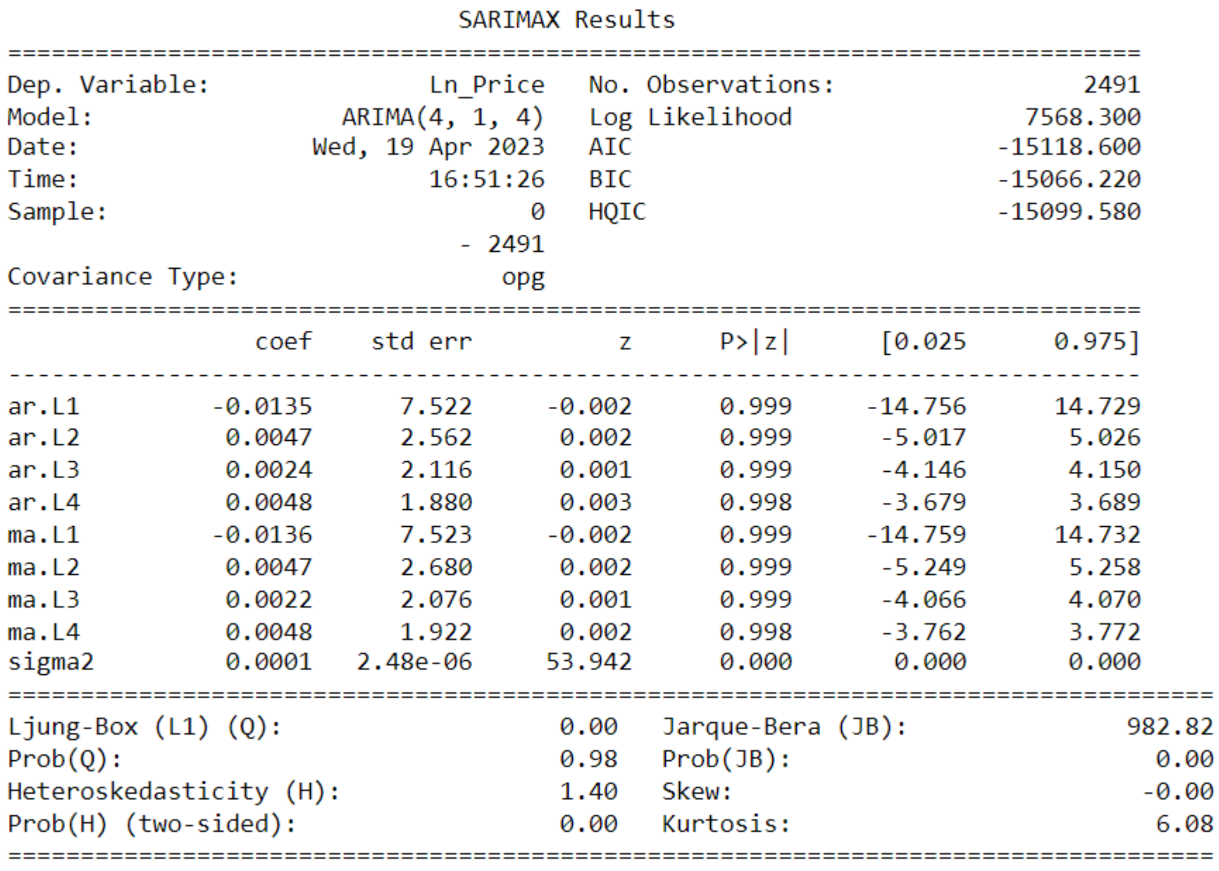
Every test set projection requires the ARIMA(p,d,q) model to be re-fit 491 times on 2000 observations. Thus, the overall process of running test set projections for 16 combinations of hyperparameters is a time-consuming process. To speed up the computations we introduce parallel processing. The following table shows the RMSE for each of the 16 combinations.

Order (p,d,q)	**RMSE**
(1, 1, 1)	0.01228
(1, 1, 2)	0.01229
(1, 1, 3)	0.01232
(1, 1, 4)	0.01230
(2, 1, 1)	0.01229
(2, 1, 2)	0.01229
(2, 1, 3)	0.01229
(2, 1, 4)	0.01237
(3, 1, 1)	0.01229
(3, 1, 2)	0.01229
(3, 1, 3)	0.01230
(3, 1, 4)	0.01229
(4, 1, 1)	0.01230
(4, 1, 2)	0.01230
(4, 1, 3)	0.01239
(4, 1, 4)	0.01226

Table 2: Hyperparameter Combination-wise RMSE

The RMSE score is least for ARIMA(4,1,4), consequently that is the best model. We now fit the ARIMA(4,1,4) model.

Table 3: ARIMA(4,1,4) model summary



```
                               SARIMAX Results
==============================================================================
Dep. Variable:               Ln_Price   No. Observations:                 2491
Model:                 ARIMA(4, 1, 4)   Log Likelihood                7568.300
Date:                Wed, 19 Apr 2023   AIC                         -15118.600
Time:                        16:51:26   BIC                         -15066.220
Sample:                             0   HQIC                        -15099.580
                               - 2491
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0135      7.522     -0.002      0.999     -14.756      14.729
ar.L2          0.0047      2.562      0.002      0.999      -5.017       5.026
ar.L3          0.0024      2.116      0.001      0.999      -4.146       4.150
ar.L4          0.0048      1.880      0.003      0.998      -3.679       3.689
ma.L1         -0.0136      7.523     -0.002      0.999     -14.759      14.732
ma.L2          0.0047      2.680      0.002      0.999      -5.249       5.258
ma.L3          0.0022      2.076      0.001      0.999      -4.066       4.070
ma.L4          0.0048      1.922      0.002      0.998      -3.762       3.772
sigma2         0.0001   2.48e-06     53.942      0.000       0.000       0.000
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               982.82
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               1.40   Skew:                            -0.00
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.08
===================================================================================
```

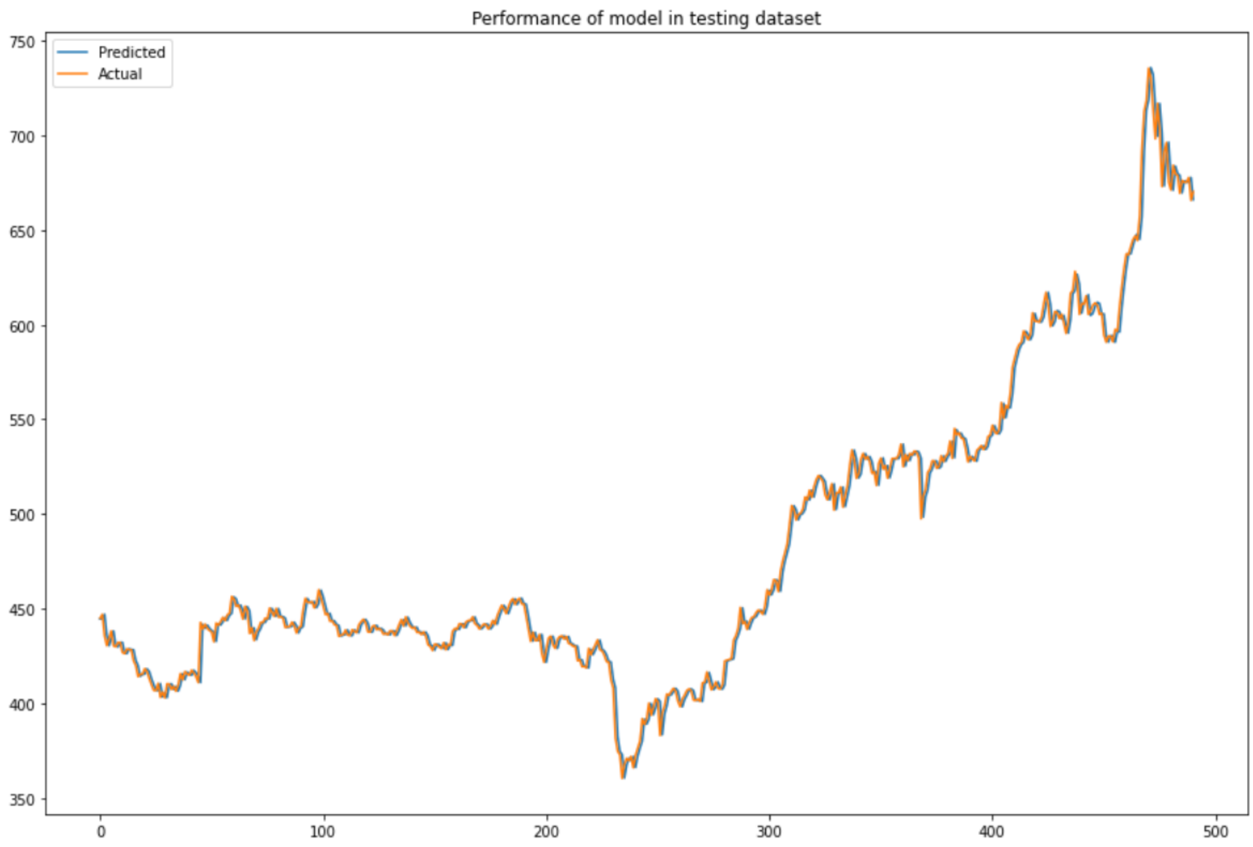


Figure 5: ARIMA(4,1,4) model performance in test dataset

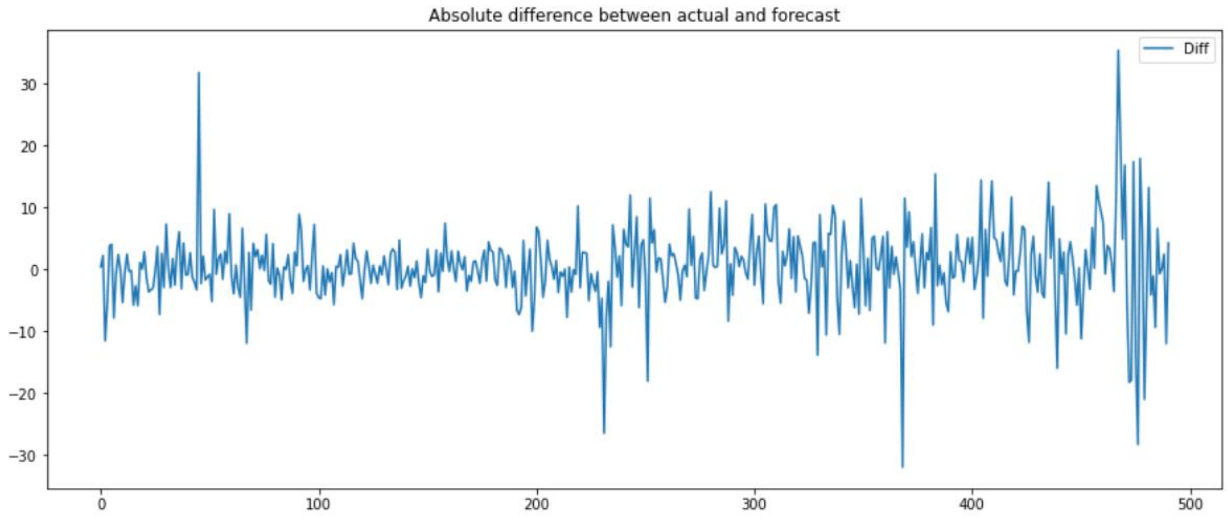


Figure 6: Absolute difference between actual and predicted values