TIME SERIES ANALYSIS FOR ETHEREUM PRICE

2021-2022

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

INSTRUCTOR:

Dr. AMIT MITRA

COMPLETED BY:

SHWETANK SINGH - 201422

SANGITA RATHOD-201402

NISTHA SHAH -201358

NIKHIL MUNAKHIYA-201352

GADDAM PRIYANKA-201311

Table of content

1. Introduction
2. Data description
   2.1 Source of data
3. Theory
   3.1 Relative Ordering Test
   3.2 Turning point test
   3.3 Augmented Dickey-Fuller Test
   3.4 ARMA Model
   3.5 ACF and PACF plot
4. Data at glance
   4.1 EDA
   4.2 Testing and Elimination of Deterministic components
5. Result
6. Conclusion

Introduction

- The Cryptocurrency market is very unpredictable, any geopolitical change can impact the trend of Cryptocurrencies in the crypto market. This is why it has always drawn the interest and attention of various professionals in the field of science or commerce. In order to get ahead of the market, statisticians have always tried to analyse and forecast the Cryptocurrency price that in essence, reflects all known and unknown information in the public domain.

- This project work is an attempt to perform *Time Series Analysis of the Ethereum Cryptocurrency Price*. The motive of this project is to analyse the Ethereum Cryptocurrency closing price movement and possibly fit a suitable model to make forecasts.

- **Time series** is a collection of data points indexed over time. Time series are analysed in order to understand the underlying structure that produce the observation. Time series data is of two types:

1. Univariate Time Series - It is a time series in which observations are sequentially recorded on a single variable over time.
2. Multivariate Time Series - It is a time series in which observations are sequentially recorded on more than one variable over time.

What is Ethereum?

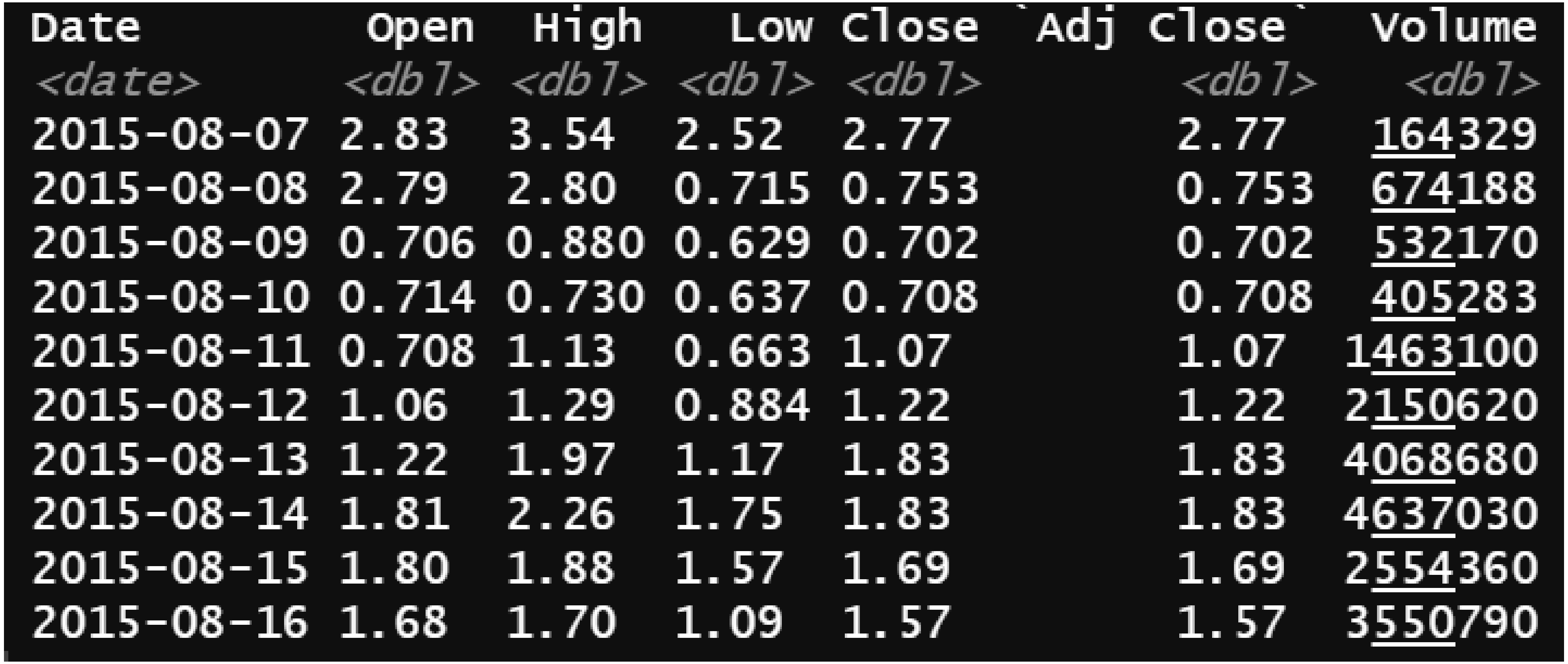
- Ethereum is open access to digital money and data-friendly services for everyone – no matter your background or location. It's a community-built technology behind the cryptocurrency ether (ETH) and thousands of applications you can use today.
- ETH is a smart contract platform that enables developers to build decentralized applications (dapps) conceptualized by Vitalik Buterin in 2013. ETH is the native currency for the Ethereum platform and also works as the transaction fees to miners on the Ethereum network. Presently ETH has second largest market capitalisation of all cryptocurrencies formed.

Data Description

- We have taken this data on Ethereum cryptocurrency over 6 years (2015 -2021). This data consists of 2221 days cryptocurrency prices (in dollars) from 01 September 2015 to 01 November 2021.
- Data can be extracted through following link:
- https://finance.yahoo.com/quote/ETH-USD/history?p=ETH-USD
- ABOUT THE DATA

1) Date – Date in format Year-Month-Day
2) Open- Opening Price at Start time
3) High- Highest Price within time window
4) Low- Low price within time window
5) Close- Close price at the end of time window
6) Volume –Volume of ETH transacted in this window

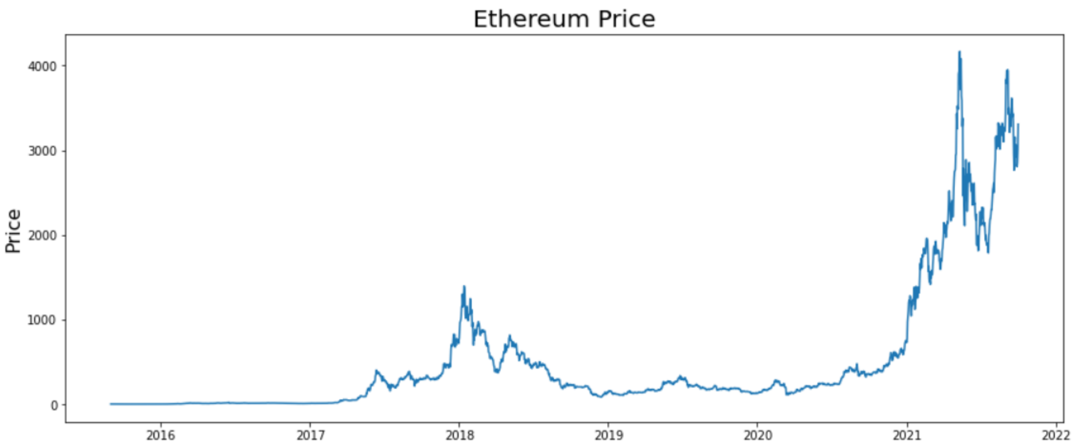
Data sample

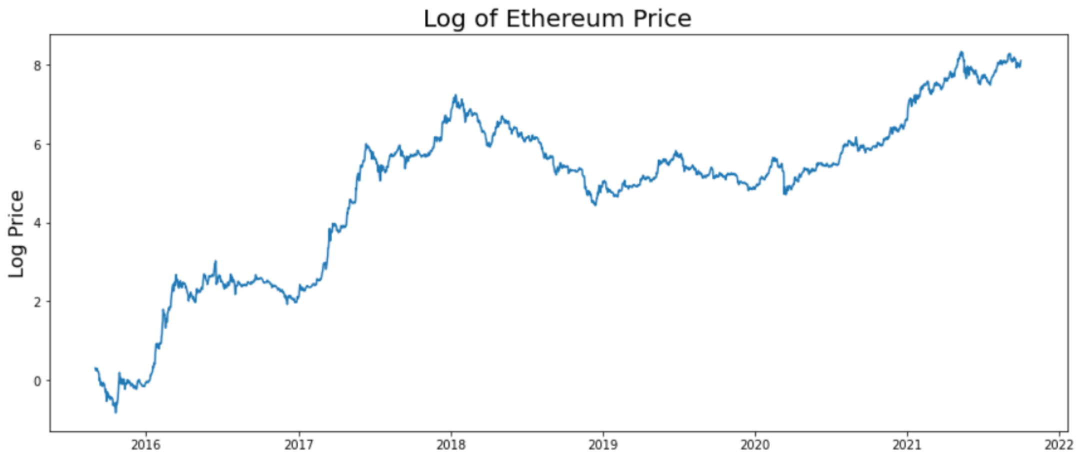


```
Date         Open  High    Low Close `Adj Close`   Volume
<date>      <dbl> <dbl> <dbl> <dbl>       <dbl>    <dbl>
2015-08-07 2.83  3.54  2.52  2.77        2.77    164329
2015-08-08 2.79  2.80  0.715 0.753       0.753   674188
2015-08-09 0.706 0.880 0.629 0.702       0.702   532170
2015-08-10 0.714 0.730 0.637 0.708       0.708   405283
2015-08-11 0.708 1.13  0.663 1.07        1.07   1463100
2015-08-12 1.06  1.29  0.884 1.22        1.22   2150620
2015-08-13 1.22  1.97  1.17  1.83        1.83   4068680
2015-08-14 1.81  2.26  1.75  1.83        1.83   4637030
2015-08-15 1.80  1.88  1.57  1.69        1.69   2554360
2015-08-16 1.68  1.70  1.09  1.57        1.57   3550790
```

Exploratory data Analysis

We plot the original data (X_t) and the log transformed data ($\log X_t$)


Ethereum Price

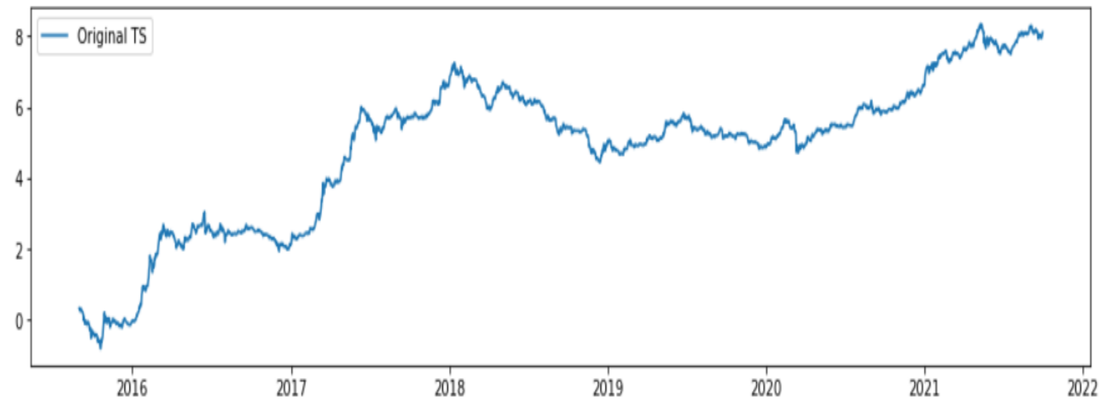

Log of Ethereum Price

The log transformed data is considered for further analysis because taking logarithm smooths the trend curve and in turn, helps to forecast easily. Thus our concerned model is given by:

$$\log X_t = Y_t = m_t + s_t + \varepsilon_t$$

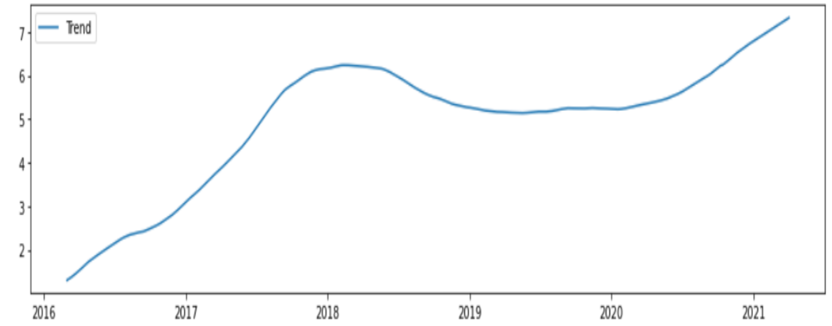
Time series Plot

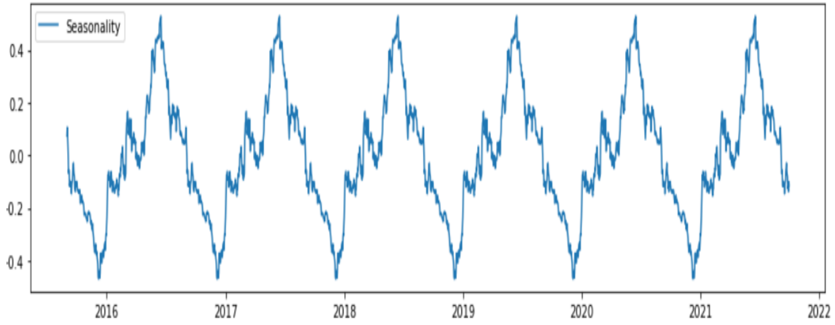
- An attempt is made to forecast the Ethereum price for the last 10 values. Keeping in mind that stock market is functional during all 365 days of the year, the train data is decomposed into its deterministic and stochastic components as shown

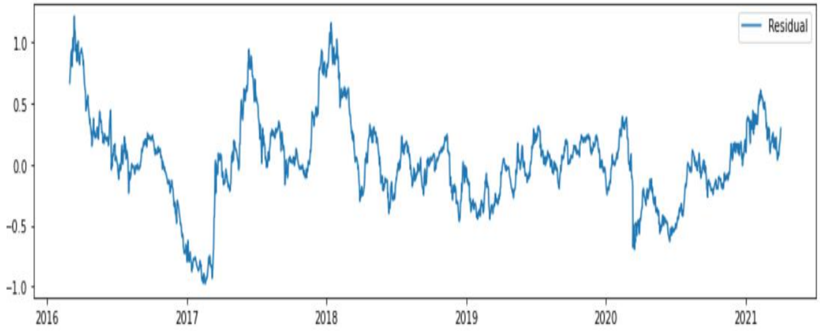


Decomposing time series data

- Time series decomposition is a process of deconstructing a time series into the following components:
- **Trend** — general movement over time
- **Seasonal** — behaviors captured in individual seasonal periods
- **Residual** — everything not captured by trend and seasonal components
- We can see that there is an increasing trend in the data. Hence we use relative ordering test.

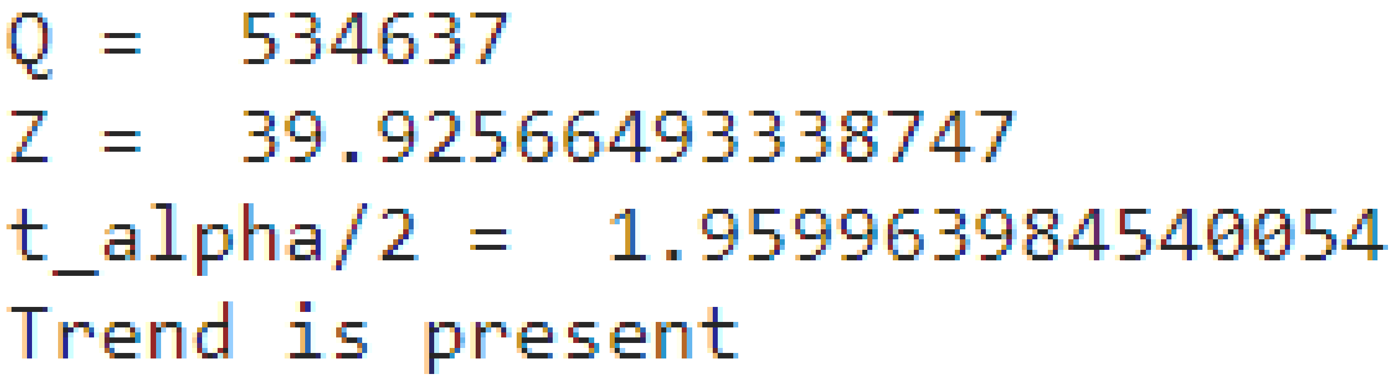






RELATIVE ORDERING TEST

- This is a non-parametric test procedure used for testing the existence of trend components.
- Null Hypothesis H_0: There is no trend in the time series
  Against
- Alternate Hypothesis H_1: There is a trend in the time series
- If observed the Q << E(Q) then it would be an indication of rising trend and if observed Q>> E(Q) then it would be the indication of falling trend. If the observed Q does not differ “significantly” from E(Q) (Under the null hypothesis) then it would indicate no trend. Q is related to the Kendall’s τ the rank correlation coefficient through the relationship
- Test Statistics: $Z = \frac{\tau - E(\tau)}{\sqrt{v(\tau)}}$) follows N(0,1) asymptotically (under Null hypothesis) We would reject the null hypothesis of no trend at level of significant α if observed |z| > $\tau_{\alpha/2}$ where $\tau_{\alpha/2}$ is the α/2th upper cut off points of a standard normal distribution.



```
Q =  534637
Z =  39.92566493338747
t_alpha/2 =  1.959963984540054
Trend is present
```

RELATIVE ORDERING TEST

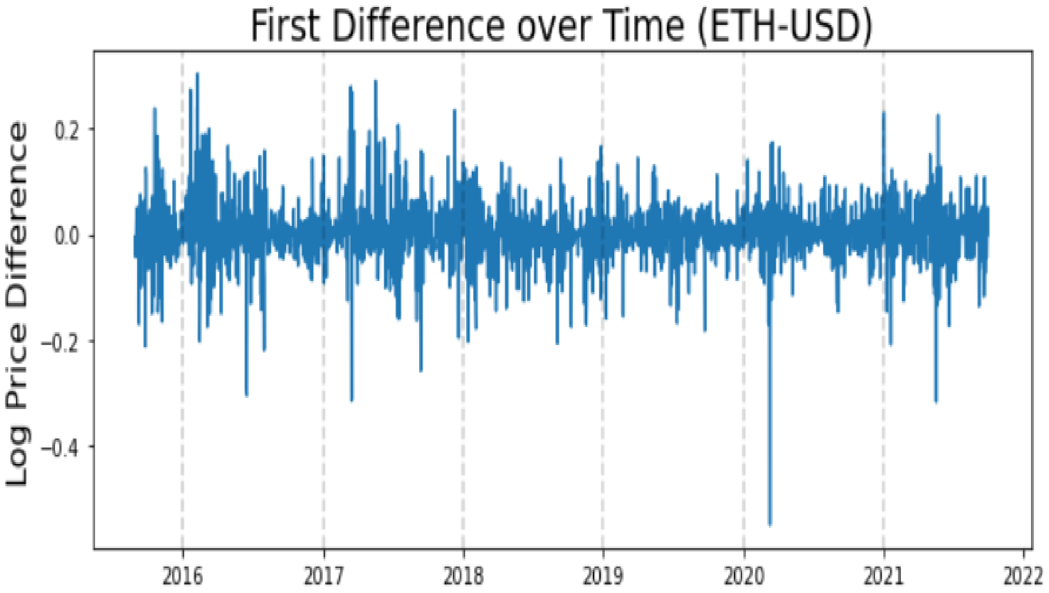
- We found that there was a significant trend present. The value of test-statistic came out be 39.9256 which is quite off from the critical z–values which was also visible from the graph

Testing and elimination of Deterministic component

- Then in order to remove trend, we applied a differencing operator of lag 1 on our data. The resultant series Z_t was obtained by the following relation,

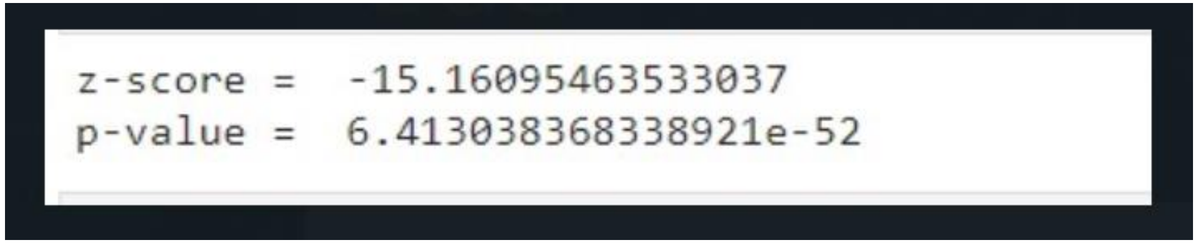
$$Z_t = Y_t - Y_{t-1}$$

- Looking at the graph, we then hypothesized that our de-trended series is purely random and free from any deterministic fluctuations
- For testing this hypothesis, we applied the Turning point test (discussed in the next slide).The evidence from the data was insufficient and we failed to reject the Null hypothesis that the series is purely random. So now we have a series which has no deterministic components in it. A natural way to proceed will be to fit standard time series models on it and see which gives the best approximation. However, before doing that, we need to first verify that the series is stationary. For that, we applied the Augmented Dickey-Fuller Test



Turning Point Test

- Turning point test is a non-parametric test which is used for testing randomness of the time series data set. Let $X_1, X_2, X_3, . . . X_n$ be the data, then X_i is considered as a turning point if either $X_{i-1} < X_i$ and $X_i > X_{i+1}$ or $X_{i-1} > X_i$ and $X_i < X_{i+1}$. We count the number of turning points in the data.
- Null Hypothesis H_0: Series is truly random (Does not contain any deterministic components.)
  Against the
  Alternative against H_1: Series is not truly random.
- Test statistics: $Z = \frac{P - E(P)}{\sqrt{\frac{16n-29}{90}}} = \frac{P - \frac{2(n-2)}{3}}{\sqrt{\frac{16n-29}{90}}}$ follows N(0, 1) asymptotically. We would reject null hypothesis H_0 at level of significance α if observed |Z| > $\tau_{\alpha/2}$ Where $\tau_{\alpha/2}$ is upper α/2 cut off point of N(0,1).



```
z-score =  -15.1609546353337
p-value =  6.413038368338921e-52
```

- So we fail to reject the null Hypothesis that the series is purely random.

Augmented Dickey-Fuller test

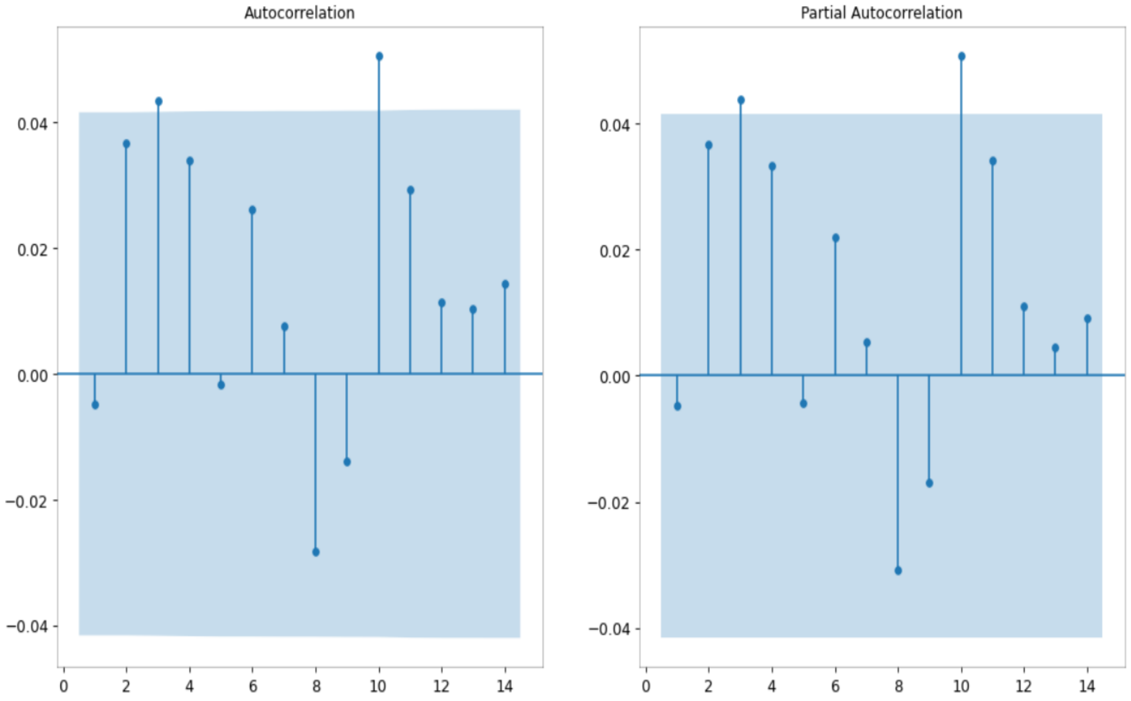
- An Augmented Dickey–Fuller test (ADF) is a test for a unit root in a time series sample. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.
- The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.
- H_0 : Data is non stationary or $\varphi_1{}^{*}= 0$
- H_1 : Data is Stationary or $\varphi_1{}^{*} < 0$
- Test statistic , $\tau^{\Lambda}{}_{\mu} = \varphi_1{}^{*} / \mathrm{SE}(\varphi_1{}^{*})$
- We reject H_0 if $\tau^{\Lambda}{}_{\mu} <$ DFα. From the test we have found out that $\tau^{\Lambda}{}_{\mu} = -8.6969$ and
- Result
- DF0.05 = −2.863, so, we reject the null at 5% level, i.e., the residual data we have derived is **stationary one.**
- We rejected the null hypothesis that the series is non-stationary and proceeded to fitting different models to this series.

ACF and PACF plot

- Autocorrelation and partial autocorrelation plots are heavily used in time series analysis and forecasting.
- These are plots that graphically summarize the strength of a relationship between an observation of a time series and observations at prior time steps. Plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) give us different viewpoints of time series.
- ACF describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information.
- PACF only describes the direct relationship between an observation and its lag.
- The lags p and q of AR and MA model can be determined from ACF and PACF plots.
- The first lag from the ACF plot will be considered for the MA parameter while that from the PACF plot will be considered for the AR parameter.

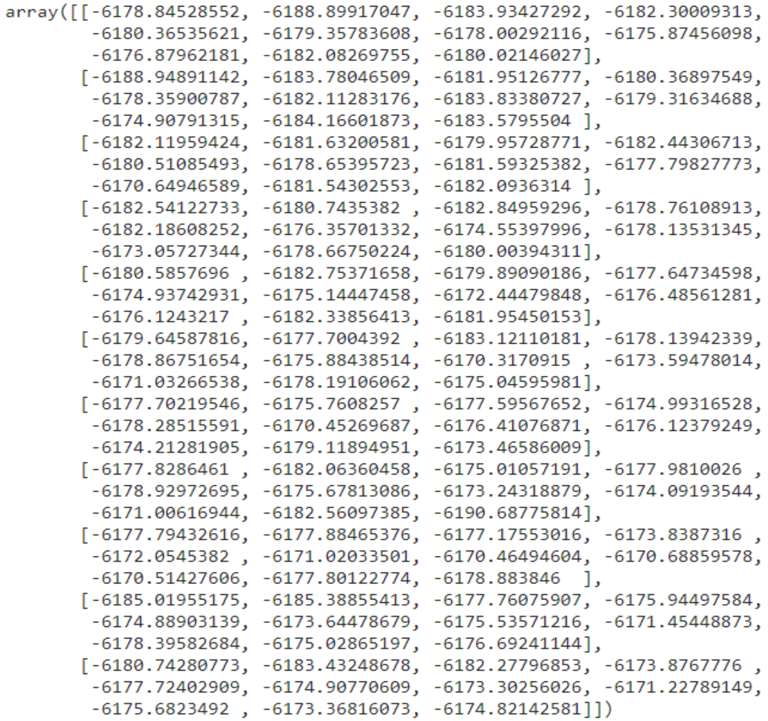
ACF and PACF plot

- For plotting ACF, we compute correlation between X_t and X_{t-k} for different lag values k and plot them in the graph. So, it is quite natural to have negative values as well. Now if the correlation values come within the significance band, then we can assume that the correlations are indifferent from zero. So, generally we take that value of k for MA as q, for which the correlation will cross the significant band for the last time i.e., we can assume current time point Xt is directly and indirectly dependent on previous q many time points. Now for plotting PACF, we regress current data point on previous time series data points.
- Then we plot the coefficients on the graph where each coefficients indicate the effect of corresponding previous data points. Now similar to ACF plot, if a value goes outside the band for some k, we assume that the observation with lag k has a direct effect on the current observation.

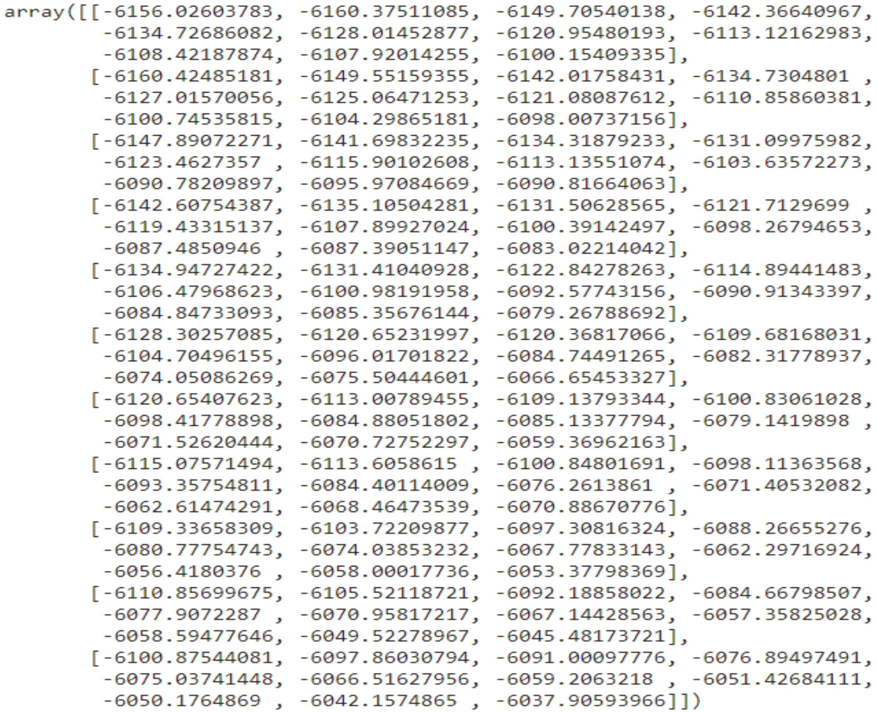


AIC and BIC values

- We will consider all the models with p and q. This gives us a total of 121 different models. From these 121 models, we will then identify the best model by the criterion of minimizing AIC and BIC. Below we have AIC and BIC values for these models.
- From these images, we conclude that the best model is ARMA(8,11) since it gives us the lowest AIC and BIC. Using ARIMA(8,1,11) on the original data we forecast the 10 future values of the Ethereum Cryptocurrency price

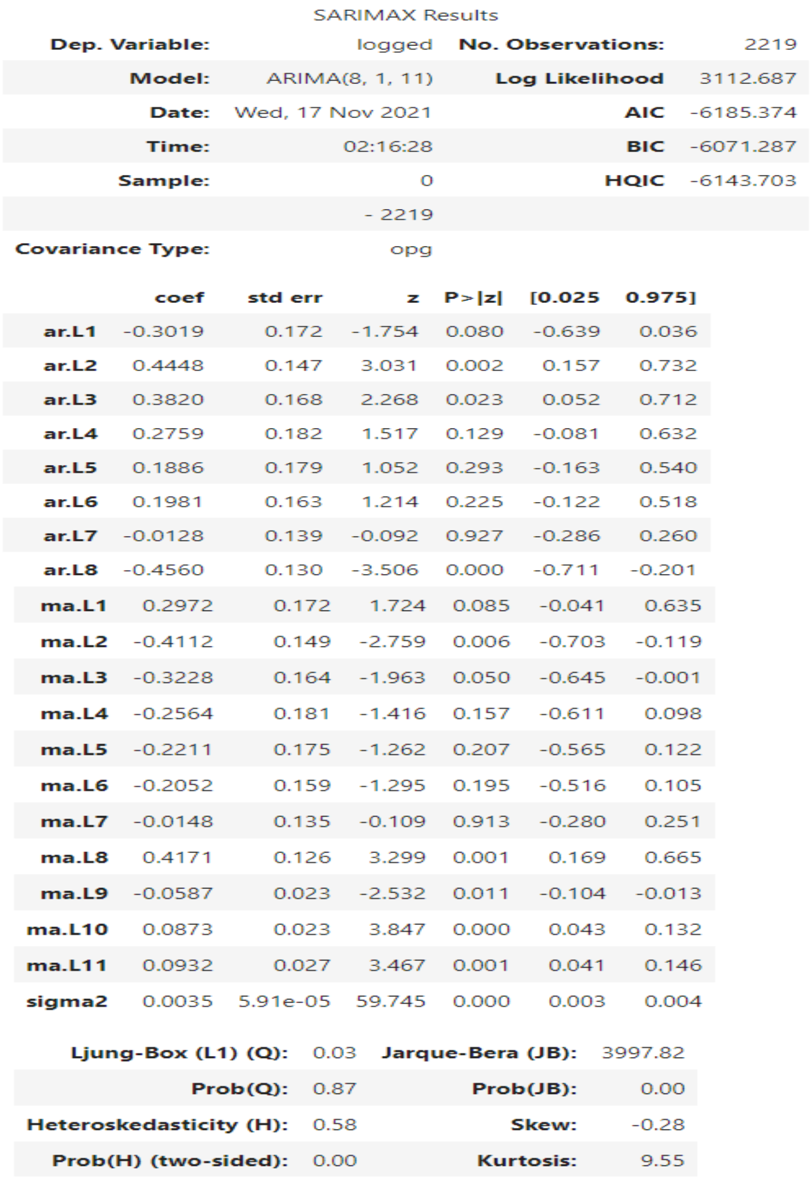


```
array([[-6178.84528552, -6188.89917047, -6183.93427292, -6182.30009313,
        -6180.36535621, -6179.35783608, -6178.00292116, -6175.87456098,
        -6176.87962181, -6182.08269755, -6180.02146027],
       [-6188.94891142, -6183.78046509, -6181.95126777, -6180.36897549,
        -6178.35900787, -6182.11283176, -6183.83380727, -6179.31634688,
        -6174.90791315, -6184.16601873, -6183.5795504 ],
       [-6182.11959424, -6181.63200581, -6179.95728771, -6182.44306713,
        -6180.51085493, -6178.65395723, -6181.59325382, -6177.79827773,
        -6170.64946589, -6181.54302553, -6182.0936314 ],
       [-6182.54122733, -6180.7435382 , -6182.84959296, -6178.76108913,
        -6182.18608252, -6176.35701332, -6174.55397996, -6178.13531345,
        -6173.05727344, -6178.66750224, -6180.00394311],
       [-6180.5857696 , -6182.75371658, -6179.89090186, -6177.64734598,
        -6174.93742931, -6175.14447458, -6172.44479848, -6176.48561281,
        -6176.1243217 , -6182.33856413, -6181.95450153],
       [-6179.64587816, -6177.7004392 , -6183.12110181, -6178.13942339,
        -6178.86751654, -6175.88438514, -6170.3170915 , -6173.59478014,
        -6171.03266538, -6178.19106062, -6175.04595981],
       [-6177.70219546, -6175.7608257 , -6177.59567652, -6174.99316528,
        -6178.28515591, -6170.45269687, -6176.41076871, -6176.12379249,
        -6174.21281905, -6179.11894951, -6173.46586009],
       [-6177.8286461 , -6182.06360458, -6175.01057191, -6177.9810026 ,
        -6178.92972695, -6175.67813086, -6173.24318879, -6174.09193544,
        -6171.00616944, -6182.56097385, -6190.68775814],
       [-6177.79432616, -6177.88465376, -6177.17553016, -6173.8387316 ,
        -6172.0545382 , -6171.02033501, -6170.46494604, -6170.68859578,
        -6170.51427606, -6177.80122774, -6178.883846  ],
       [-6185.01955175, -6185.38855413, -6177.76075907, -6175.94497584,
        -6174.88903139, -6173.64478679, -6175.53571216, -6171.45448873,
        -6178.39582684, -6175.02865197, -6176.69241144],
       [-6180.74280773, -6183.43248678, -6182.27796853, -6173.8767776 ,
        -6177.72402909, -6174.90770609, -6173.30256026, -6171.22789149,
        -6175.6823492 , -6173.36816073, -6174.82142581]])
```



```
array([[-6156.02603783, -6160.37511085, -6149.70540138, -6142.36640967,
        -6134.72686082, -6128.01452877, -6120.95480193, -6113.12162983,
        -6108.42187874, -6107.92014255, -6100.15409335],
       [-6160.42485181, -6149.55159355, -6142.01758431, -6134.7304801 ,
        -6127.01570056, -6125.06471253, -6121.08087612, -6110.85860381,
        -6100.74535815, -6104.29865181, -6098.00737156],
       [-6147.89072271, -6141.69832235, -6134.31879233, -6131.09975982,
        -6123.4627357 , -6115.90102608, -6113.13551074, -6103.63572273,
        -6090.78209897, -6095.97084669, -6090.81664063],
       [-6142.60754387, -6135.10504281, -6131.50628565, -6121.7129699 ,
        -6119.43315137, -6107.89927024, -6100.39142497, -6098.26794653,
        -6087.4850946 , -6087.39051147, -6083.02214042],
       [-6134.94727422, -6131.41040928, -6122.84278263, -6114.89441483,
        -6106.47968623, -6100.98191958, -6092.57743156, -6090.91343397,
        -6084.84733093, -6085.35676144, -6079.26788692],
       [-6128.30257085, -6120.65231997, -6120.36817066, -6109.68168031,
        -6104.70496155, -6096.01701822, -6084.74491265, -6082.31778937,
        -6074.05086269, -6075.50444601, -6066.65453327],
       [-6120.65407623, -6113.00789455, -6109.13793344, -6100.83061028,
        -6098.41778898, -6084.88051802, -6085.13377794, -6079.1419898 ,
        -6071.52620444, -6070.72752297, -6059.36962163],
       [-6115.07571494, -6113.6058615 , -6100.84801691, -6098.11363568,
        -6093.35754811, -6084.40114009, -6076.2613861 , -6071.40532082,
        -6062.61474291, -6068.46473539, -6070.88670776],
       [-6109.33658309, -6103.72209877, -6097.30816324, -6088.26655276,
        -6080.77754743, -6074.03853232, -6067.77833143, -6062.29716924,
        -6056.4180376 , -6058.00017736, -6053.37798369],
       [-6110.85699675, -6105.52118721, -6092.18858022, -6084.66798507,
        -6077.9072287 , -6070.95817217, -6067.14428563, -6057.35825028,
        -6058.59477646, -6049.52278967, -6045.48173721],
       [-6100.87544081, -6097.86030794, -6091.00097776, -6076.89497491,
        -6075.03741448, -6066.51627956, -6059.2063218 , -6051.42684111,
        -6050.1764869 , -6042.1574865 , -6037.90593966]])
```

Summary



SARIMAX Results

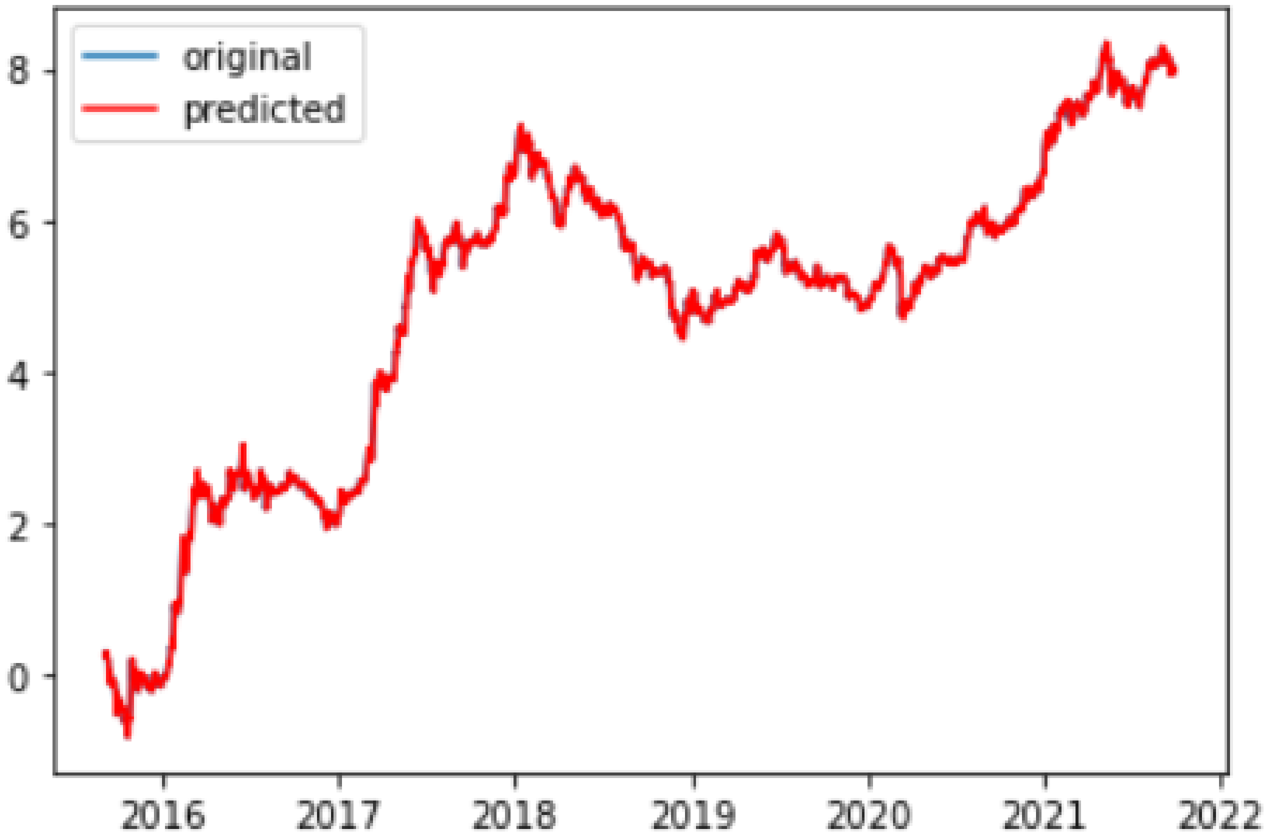
Dep. Variable:	logged	**No. Observations:**	2219
Model:	ARIMA(8, 1, 11)	**Log Likelihood**	3112.687
Date:	Wed, 17 Nov 2021	**AIC**	-6185.374
Time:	02:16:28	**BIC**	-6071.287
Sample:	0	**HQIC**	-6143.703
	- 2219		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	-0.3019	0.172	-1.754	0.080	-0.639	0.036
ar.L2	0.4448	0.147	3.031	0.002	0.157	0.732
ar.L3	0.3820	0.168	2.268	0.023	0.052	0.712
ar.L4	0.2759	0.182	1.517	0.129	-0.081	0.632
ar.L5	0.1886	0.179	1.052	0.293	-0.163	0.540
ar.L6	0.1981	0.163	1.214	0.225	-0.122	0.518
ar.L7	-0.0128	0.139	-0.092	0.927	-0.286	0.260
ar.L8	-0.4560	0.130	-3.506	0.000	-0.711	-0.201
ma.L1	0.2972	0.172	1.724	0.085	-0.041	0.635
ma.L2	-0.4112	0.149	-2.759	0.006	-0.703	-0.119
ma.L3	-0.3228	0.164	-1.963	0.050	-0.645	-0.001
ma.L4	-0.2564	0.181	-1.416	0.157	-0.611	0.098
ma.L5	-0.2211	0.175	-1.262	0.207	-0.565	0.122
ma.L6	-0.2052	0.159	-1.295	0.195	-0.516	0.105
ma.L7	-0.0148	0.135	-0.109	0.913	-0.280	0.251
ma.L8	0.4171	0.126	3.299	0.001	0.169	0.665
ma.L9	-0.0587	0.023	-2.532	0.011	-0.104	-0.013
ma.L10	0.0873	0.023	3.847	0.000	0.043	0.132
ma.L11	0.0932	0.027	3.467	0.001	0.041	0.146
sigma2	0.0035	5.91e-05	59.745	0.000	0.003	0.004

Ljung-Box (L1) (Q):	0.03	**Jarque-Bera (JB):**	3997.82
Prob(Q):	0.87	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.58	**Skew:**	-0.28
Prob(H) (two-sided):	0.00	**Kurtosis:**	9.55

Result

- We conclude that ARIMA(8,1,11) is the most suitable fit for transformed (log) prices. A plot of the fitted values superimposed over original values is shown



Final model

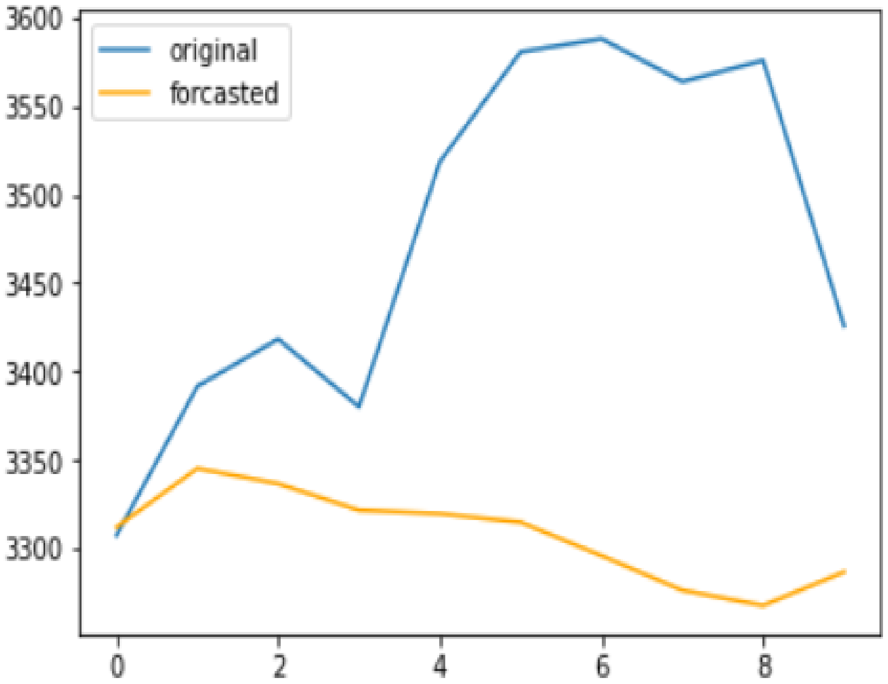
- The AIC of this model came out to be around -6190.6877 and BIC around -6070.8867. Figure 5.2 shows summary of the in-built fit function of a python library. From the summary, we see that all the coefficients of the model are significant at 5% level of significance. If the original series is $\{X_t\}$, the final model equation is written as follows,

$$Z_t = \mu + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \varphi_3 Z_{t-3} + \varphi_4 Z_{t-4} + \varphi_5 Z_{t-5} + \varphi_6 Z_{t-6} + \varphi_7 Z_{t-7} + \varphi_8 Z_{t-8} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \theta_4 \epsilon_{t-4} + \theta_5 \epsilon_{t-5} \theta_1 + \theta_6 \epsilon_{t-6} + \theta_7 \epsilon_{t-7} \theta_1 + \theta_8 \epsilon_{t-8} + \theta_9 \epsilon_{t-9} + \theta_{10} \epsilon_{t-10} + \theta_{11} \epsilon_{t-11} + \epsilon_t$$

where $Z_t = X_t - X_{t-1}$ and $\epsilon_t \sim N(0, \sigma 2)$; $\sigma^2 = 0.0035$ for all t.

Original vs forecasted values

- After forecasting the future 10 values, we compared them to the actual 10 test values. The plot for the forecasted and the actual values is shown



Conclusion

- To conclude, we found that ARMA(8,11) fits best on the series obtained after first order differencing i.e., ARIMA(8,1,11) fits best for the original data on Ethereum Cryptocurrency prices. The model equation is given by (1).
- We employed the standard approach of first eliminating the trend from the data. We used first order differencing for this. After eliminating the trend, we found that the resultant series was purely random and stationary. We plotted the ACF and PACF plots for this resultant series and identified the candidate models for this data. Finally, we fitted all the models and based on the criterion of minimizing AIC, we found the best model. ,The best model was able to fit the data quite well.