IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – VI
Auto-regression

Student's Name: Abhay Vijayvargiya **Mobile No: 6377967485**

Roll Number: B20176 **Branch:DSE**

1 a.

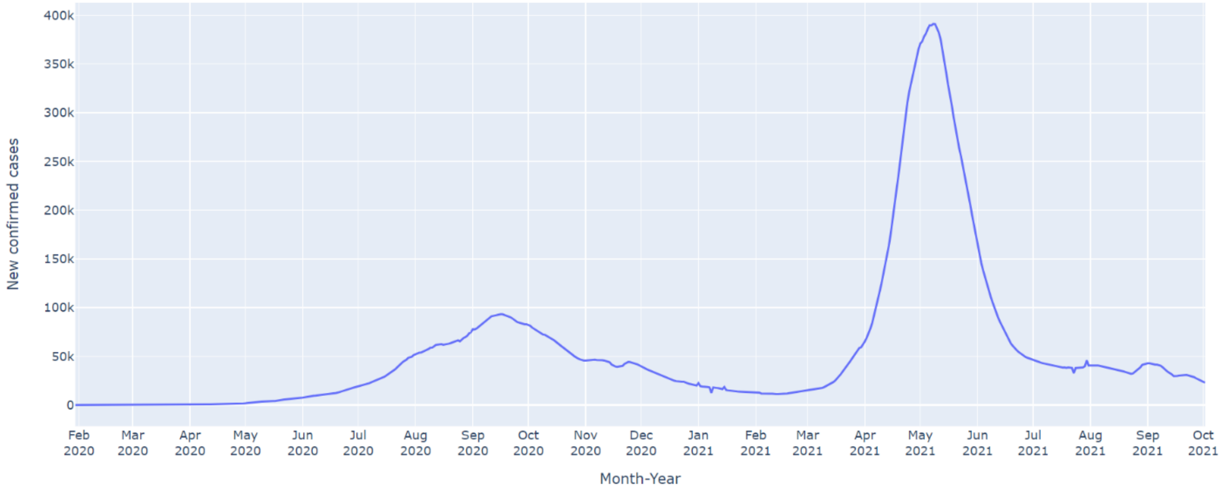


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The days one after the other have roughly the same no. of new confirmed cases.
2. The no. of new covid cases on a particular day is correlated to yesterday's no. of new covid cases.
3. First wave occurred from mid May 2020 to February 2020 and second wave occurred from March 2021 to August 2021.

b. The value of the Pearson's correlation coefficient is 0.999

Inferences:

1. From the value of Pearson's correlation coefficient, the two-time sequences are highly correlated.
2. The Pearson's correlation coefficient exactly meets our expectation of having similar observation between consecutive sequences.
3. Since the no. if new confirmed cases cannot increase or decrease rapidly in one day we can expect that the data is autocorrelated.

c.

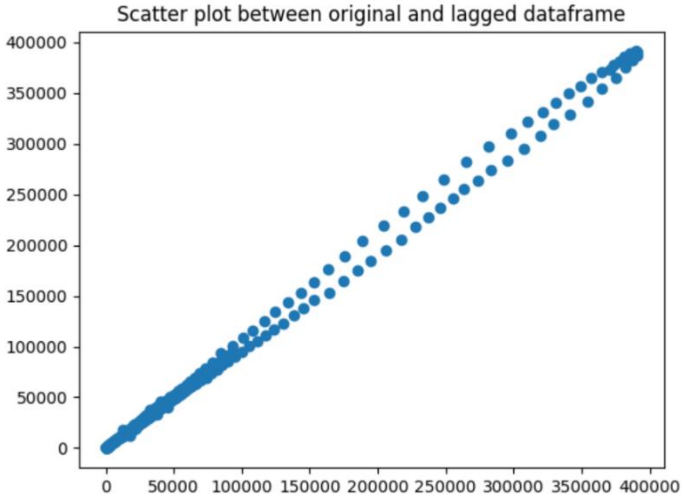


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the spread of the data points the correlation coefficient of the lagged sequence and given time sequence is very high.
2. The above scatter plot correctly signifies the correlation between the two-time sequences.
3. Since the no. if new confirmed cases cannot increase or decrease rapidly in one day, we can expect that the data is highly autocorrelated.

d.

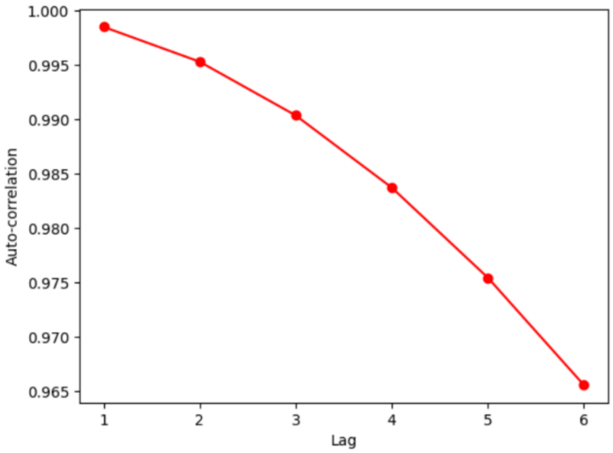


Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The value of coefficient of correlation is decreasing with the increasing time lag.
2. The coefficient is decreasing because there is less correlation among 6, 5, 4, days with the given time sequence.

e.

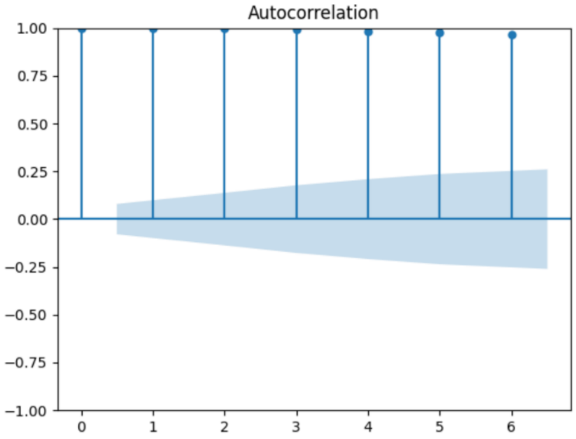


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient is continuously decreasing with increase in time lags.
2. The coefficient is decreasing because there is less correlation among 6, 5, 4, days with the given time sequence. Since the no. if new confirmed cases cannot increase or decrease rapidly in one day, we can expect that the data is highly autocorrelated.

2

a. The coefficients obtained from the AR model are; 59.95, 1.036, 0.261, 0.0275, -0.175, -0.152.

b. i.

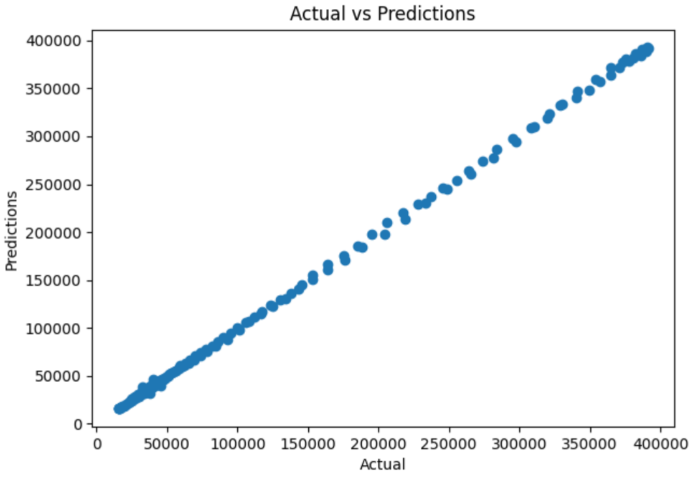
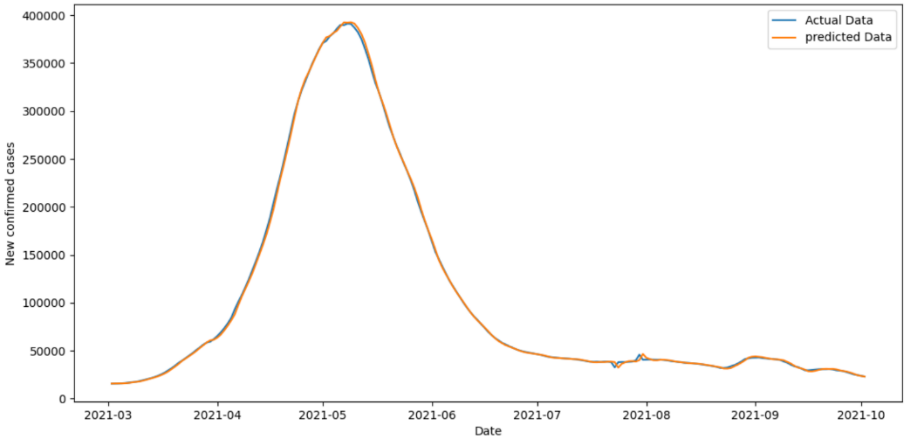


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the spread of the data points, it is observed that the predicted data is very accurate.
2. The scatter plot seems to obey the nature reflected by Pearson’s correlation coefficient calculation.
3. Since we have taken the time lags up to 5 days our predicted data is matching significantly with actual data.

ii.



Line plot between actual data and predicted data

Inferences:

1. The model is very accurate for predicting the test data since we have taken up to 5 time day-lags with very high correlation coefficient. The predicted data is overlapping the actual data.

iii.

The RMSE (%) and MAPE between predicted power consumed for test data and original values for test data are 2.66 and 1.57 respectively.

Inferences:

1. Lower RMSE and MAPE values represents higher accuracy of the model built.

3

Table 1 RMSE (%) and MAPE between predicted and original data values wrt lags in time sequence

Lag value	RMSE (%)	MAPE
1	4.38	3.44
5	2.66	1.57
10	2.52	1.52
15	2.50	1.50
25	2.51	1.53

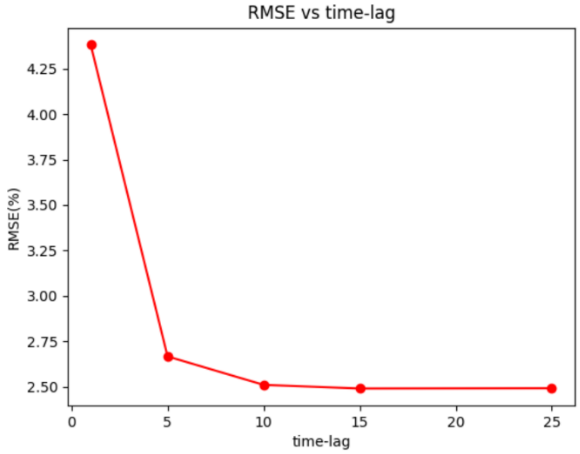


Figure 6 RMSE (%) vs. time lag

Inferences:

1. The RMSE decreases with increase in lag up to a certain point and then increases slowly.

2. Since we are taking more lag with higher correlation, we are getting low RMSE. But further increment in time lag will produce more error because we are taking account of time lags which have low correlation coefficient.

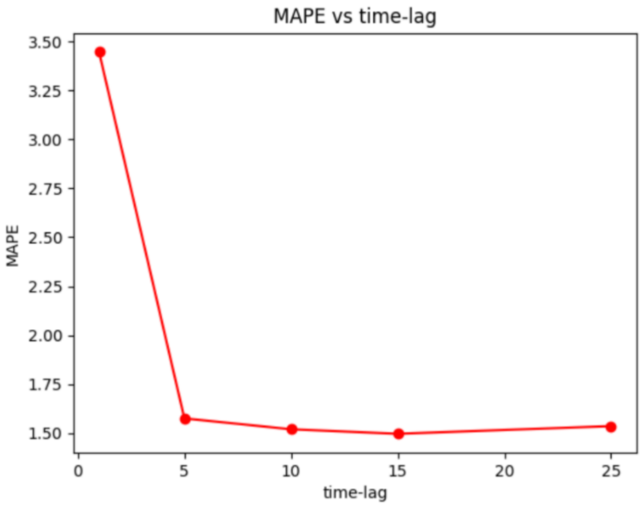


Figure 7 MAPE vs. time lag

Inferences:

1. The MAPE decreases with increase in lag up to a certain point and then increases slowly.
2. Since we are taking more lag with higher correlation, we are getting low MAPE. But further increment in time lag will produce more error because we are taking account of time lags which have low correlation coefficient.

4

The heuristic value for the optimal number of lags is 78.

The RMSE (%) and MAPE value between test data time sequence and original test data sequence are 2.95 and 2.075

Inferences:

1. No based upon the heuristic value for the lag we are getting some what high error than previously used time lags.
2. Since we have taken up to 78 lags, those which does not have good correlation are also taken in account which results in less accuracy of the model.

3. RMSE and MAPE are least for the time lag of 15 days and above this the error is increasing.