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

Student's Name: Gaurav Kumar **Mobile No: 8529143452**

Roll Number: B20197 **Branch:EE**

1 a.

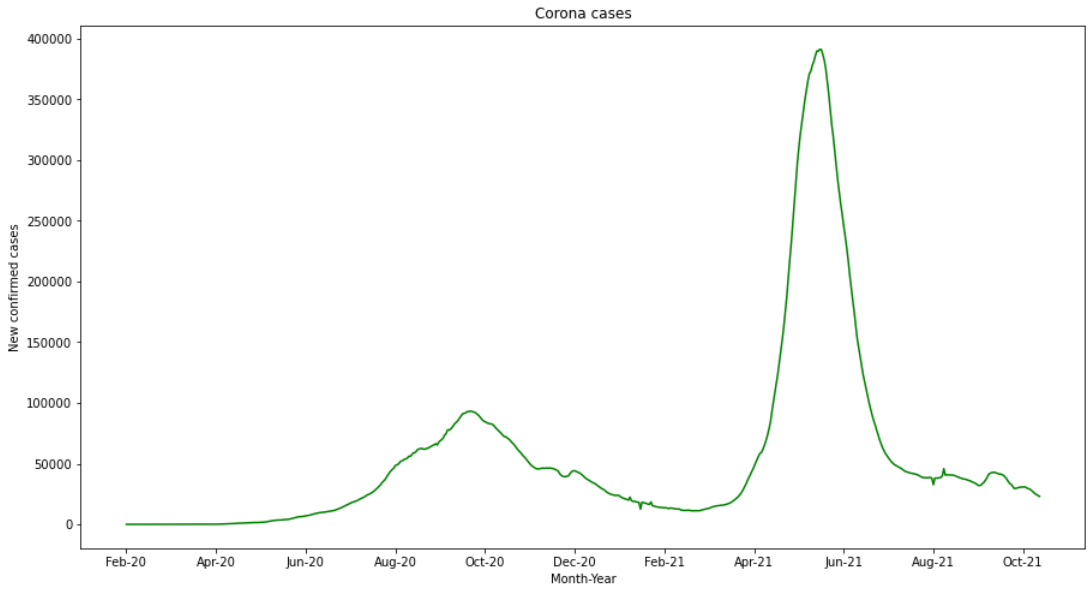

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The plot shows that the data if dependent is exponential in nature, with some seasonality.
2. The data has curves similar to exponential data for the rise and fall parts and peak is formed in both years 2020 and 2021.
3. The duration of first wave spans from July 2020 to Jan 2021 i.e., around 7 months with initial 3 months of time for increasing cases tremendously and next 4 months with decrease in new cases with high rate. The duration of second wave spans for shorter duration, April 2021 to July 2021, i.e., around 4 months with initial 2 months for tremendous increase in new cases and other 2 months for decrease in cases with higher rate.
4. More people were affected by wave 2 than by wave 1 as the peak of wave 2 is much higher than the peak of wave 1 and the duration of wave 2 is also more than the half of wave 1.

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

Inferences:

1. As the Pearson's coefficient is approximately 1, therefore, we can say that x(t) and x(t-1) are strongly correlated and x(t) highly depends on x(t-1).
2. The Pearson's coefficient value suggests that the values of cases one day after a particular day would be almost same with a very little change.
3. The data of covid cases has time dependence, the number of cases in next day depends on the cases of the present day, thus the high value of Pearson's coefficient and similar values for next day new cases. The cause for dependence may have underlying structures like the state of current distribution of covid patients, availability of hospitals, the movement pattern of covid infected etc. These underlying properties change over a few days and not suddenly so it is expected that the outcome , i.e. number of covid cases will also not change dramatically in a single day.

c.

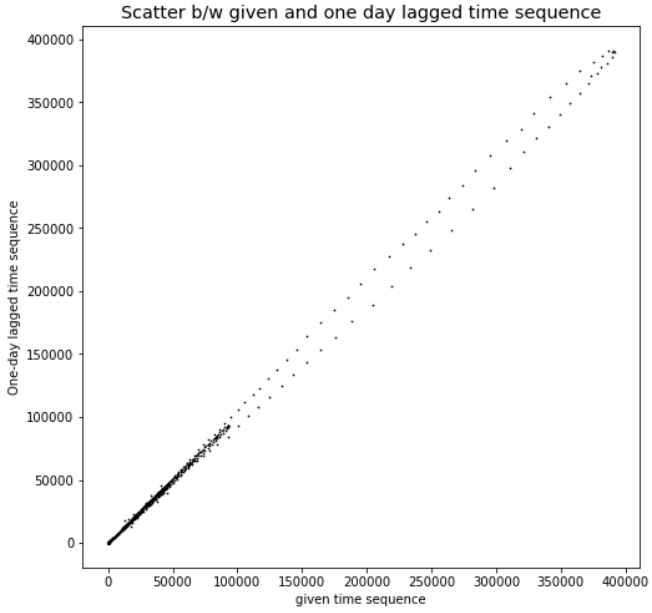

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. X(t) and X(t-1) forms an almost straight line with positive slope; thus, both the lists are strongly dependent on each other with high positive correlation.
2. The scatter plot obeys the nature reflected by Pearson's correlation coefficient.
3. Both suggest that the data is strong correlated as the Pearson's coefficient value is approximately 1 and the best fit line in scatter plot would be a line with positive slope and all the points being very close to that line.

d.

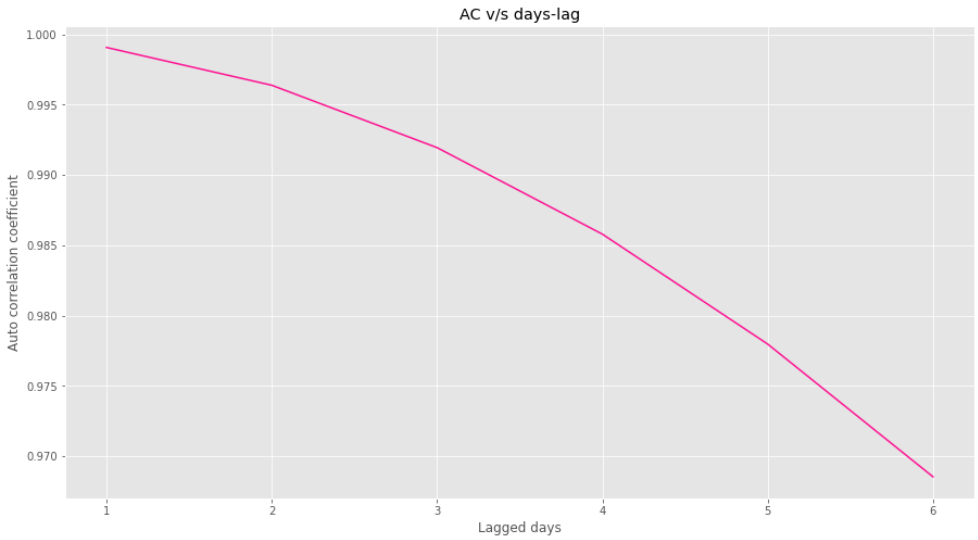

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The value for correlation coefficient decreases with increase in number of lags.
2. The data would be more dependent on the data just before it rather than the data which was a few days back, there is some correlation but the correlation would decrease with time.

e.

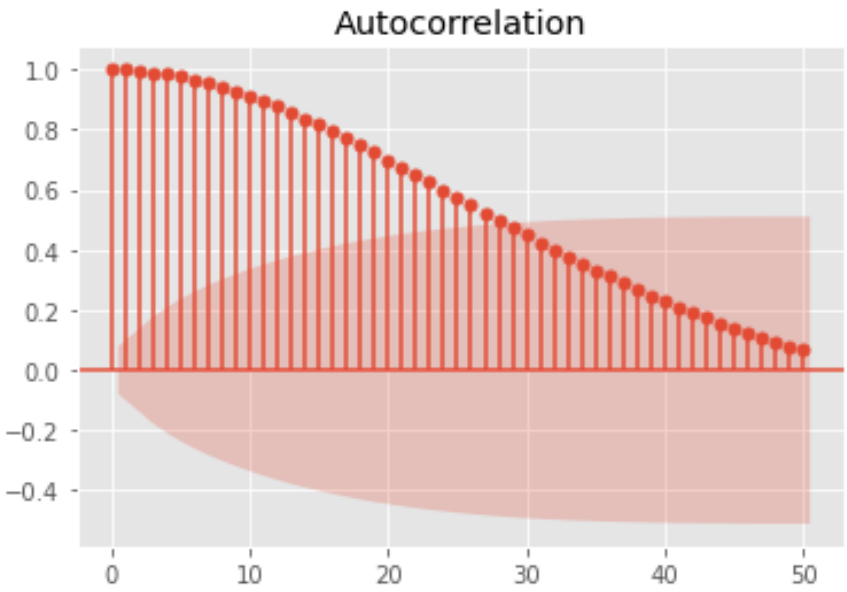

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The value of correlation coefficient decreases with increase in number of lags.
2. The data would be more dependent on the data just before it rather than the data which was a few days back, there is some correlation but the correlation would decrease with time.

2

a. The coefficients obtained from the AR model are;
Const = 59.954, L1 = 1.036, L2 = 0.261, L3 = 0.027, L4 = -0.175, L5 = -0.152

b. i.

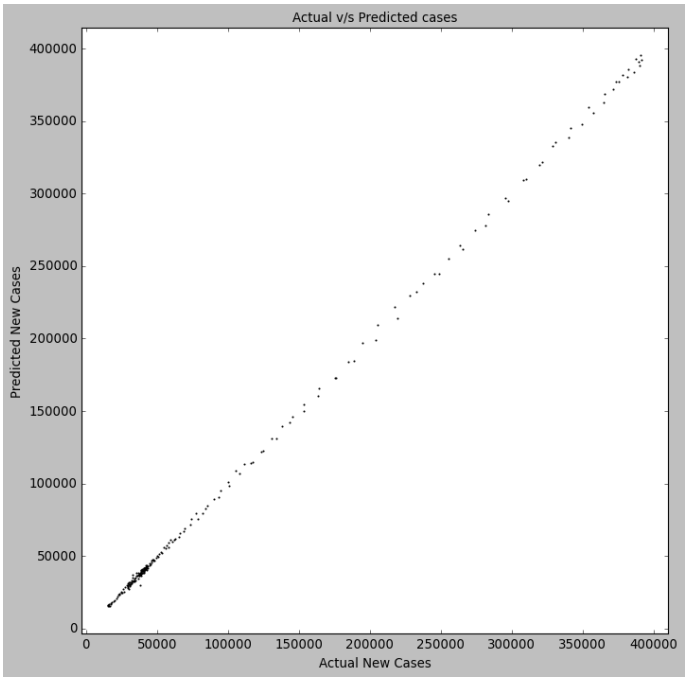

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. The scatter plot itself is like a straight line with positive slope denoting high positive correlation of predicted value and actual value.
2. The scatter plot itself is like a straight line with positive slope denoting high positive correlation of predicted value and actual value.
3. Both suggest that the data is strongly correlated as the Pearson's coefficient value is approximately 1 and the best fit line in scatter plot would be a line with positive slope and all the points being very close to that line.

ii.

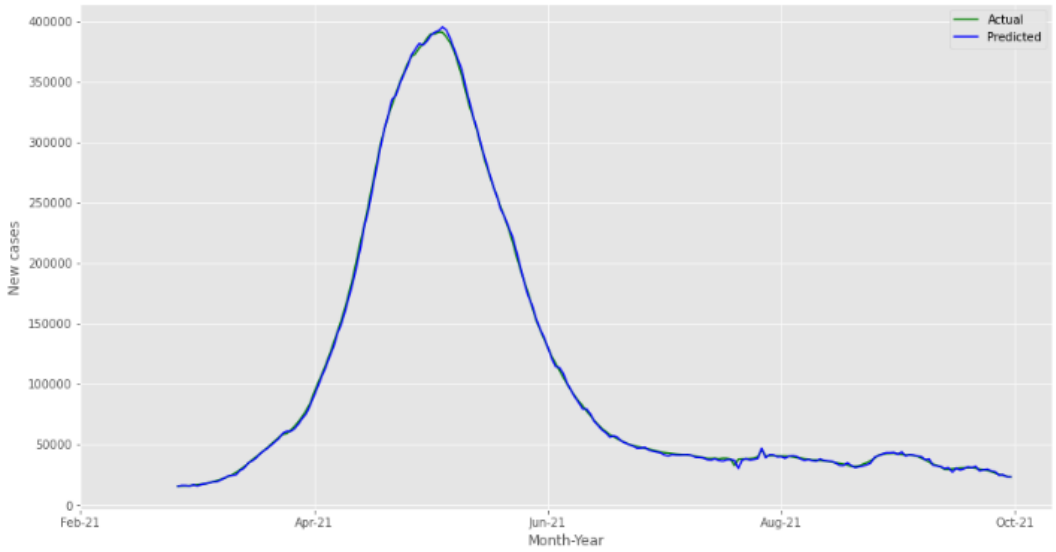

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. As the plot for predicted and actual data is almost overlapping at all the points, therefore, we can say that the model is highly reliable for future predictions

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.852 and 1.575 respectively.

Inferences:

1. As the RMSE and MAPE errors are very less (<2%) therefore we can say that the model is highly accurate and hence reliable.
2. The lower error is because the predicted value and the actual value are close to each other, which us the main requirement for the model thus the model is reliable.

3

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

Lag value	RMSE (%)	MAPE
1	5.375	3.447
5	1.825	1.575
10	1.686	1.519
15	1.612	1.496
25	1.703	1.535

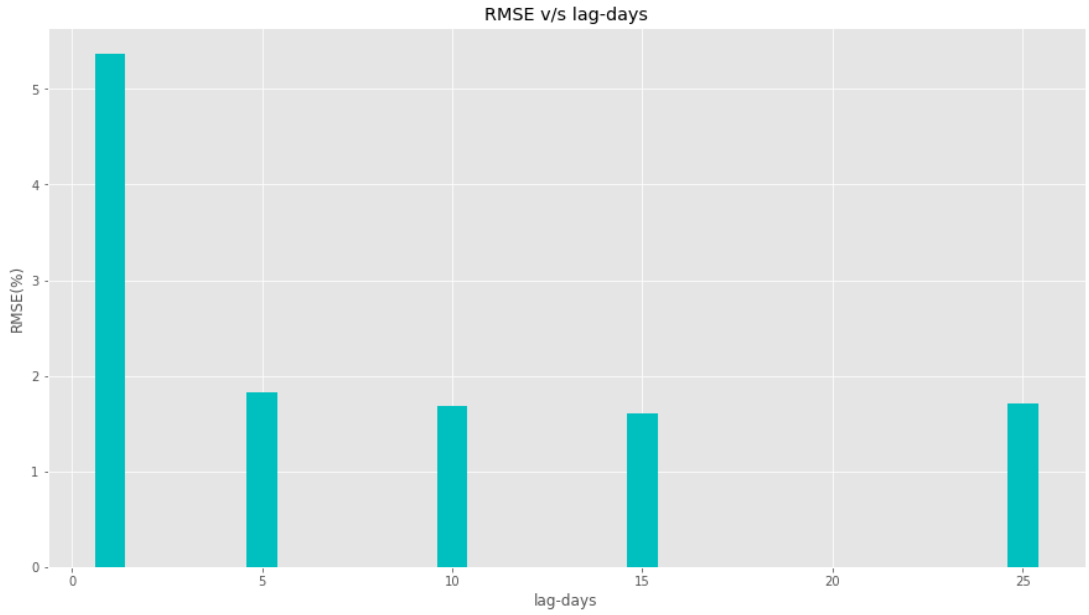

Figure 7 RMSE(%) vs. time lag

Inferences:

1. RMSE error decreases up till certain lag value (p = 15) and then starts increasing.
2. The data is dependent on past values; thus, the data depends on more lag value, thus as we increase p value, the error decreases but after a certain p value, the dependence of the value with its past values is not that dependent which leads to increase in error but that increases is small.

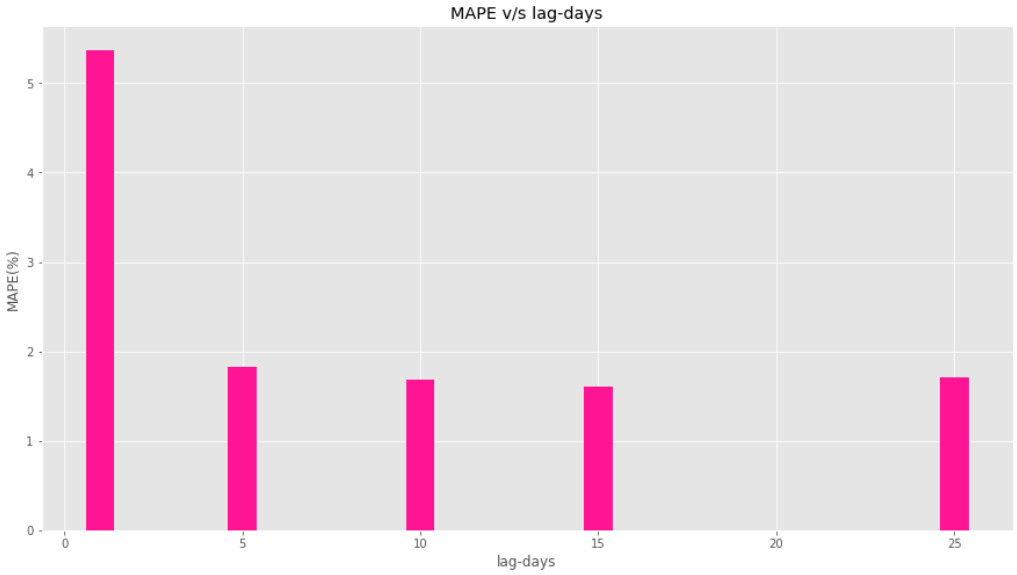

Figure 8 MAPE vs. time lag

Inferences:

1. MAPE error decreases up till certain lag value (p = 15) and then starts increasing.
2. The data is dependent on past values; thus, the data depends on more lag value, thus as we increase p value, the error decreases but after a certain p value, the dependence of the value with its past values is not that dependent which leads to increase in error but that increases is small.

4

The heuristic value for the optimal number of lags is 77

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are 1.759% and 2.026%.

Inferences:

1. Yes, heuristics for calculating the optimal number of lags improve the prediction accuracy of the model.
2. The value of RMSE and MAPE at heuristic value is more than value of RMSE and MAPE at lag value = 15, so the heuristic value does not give optimal value, but as the difference in the errors is small it can be used rather than finding the minimum error lag value which could be a bit exhaustive.
3. The RMSE values decreases up till certain p value, and then increases slowly till p value equal to heuristic value.