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

Student's Name: Yash Sharma **Mobile No: 8802131138**

Roll Number: B20241 **Branch: CSE**

1 a.

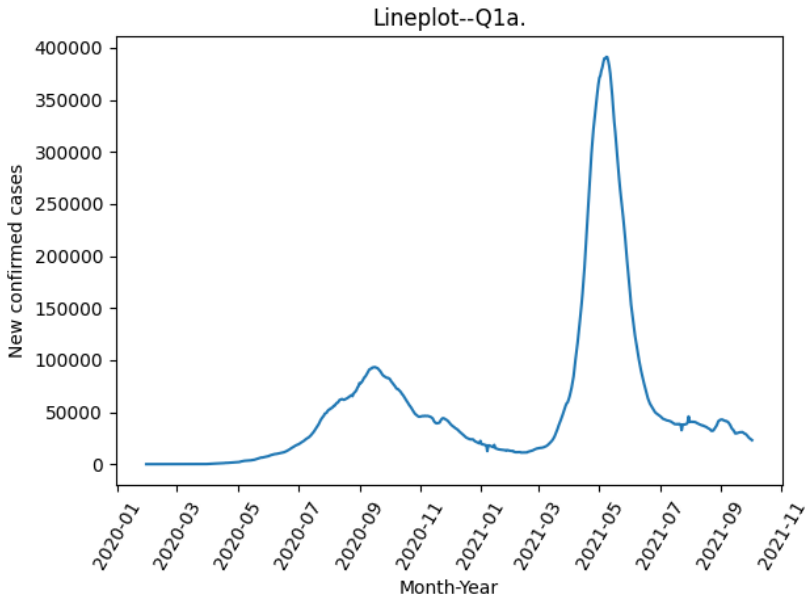

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. It may be observed that the days one after the other have quite similar values and there is no sudden change seen anywhere.
2. The reason for such observation is because in the given time series data any particular value is dependent on its previous values to some extent.
3. The first wave was from 30-07-2020 to 30-12-2020 (roughly 5 months).
4. The second wave started approximately on 30-03-2021 and ended on 31-07-21 which is which is approximately 4 months long.

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

Inferences:

1. From the observed value which nearly rounds off to unit correlation infers that the data we have shows a very strong positive correlation with its previous value (lag of 1 day).
2. There is a very high degree of correlation between the two-time sequences.
3. The observations (number of covid cases) are expected to be dependent on the previous values because for the given data, it is not expected that any sharp change in number of cases would occur over a day or night or so. The process is quite gradual and so a trend of increase or decrease in number of observed may be observed and thus the data is expected to dependent on previous values. The same can be verified when we see the very high degree of correlation between the two-time sequences.
4. The stronger the correlation between the output variable and a specific lagged variable, the more weight that autoregression model can put on that variable when modeling.
5. Thus, by observing the Pearson coefficient between the time sequences we may easily infer that the future covid cases would surely depend on the previous inputs and thus lagged time sequences of the same may be used to predict any further covid cases.

c.

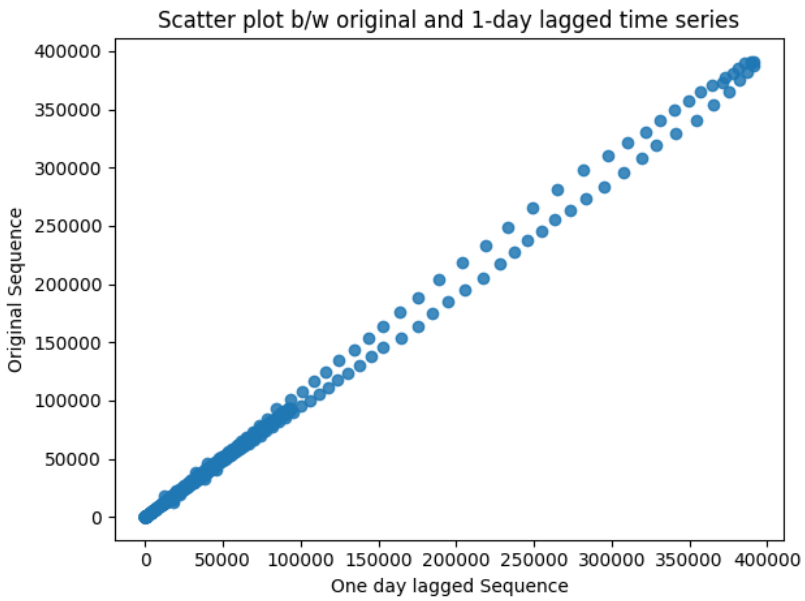

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From observing the scatter plot, we may infer that the two-time sequences are strongly positively correlated with each other.

2. The scatter plot exactly follows the nature reflected by Pearson's coefficient in the previous problem (0.999).
3. The reason for the above inference is that the given time-series data has very high dependency on the previous 1 day lagged values and that is clearly visible when we observe the Pearson coefficient between the same (which is indeed close to unity).

d.

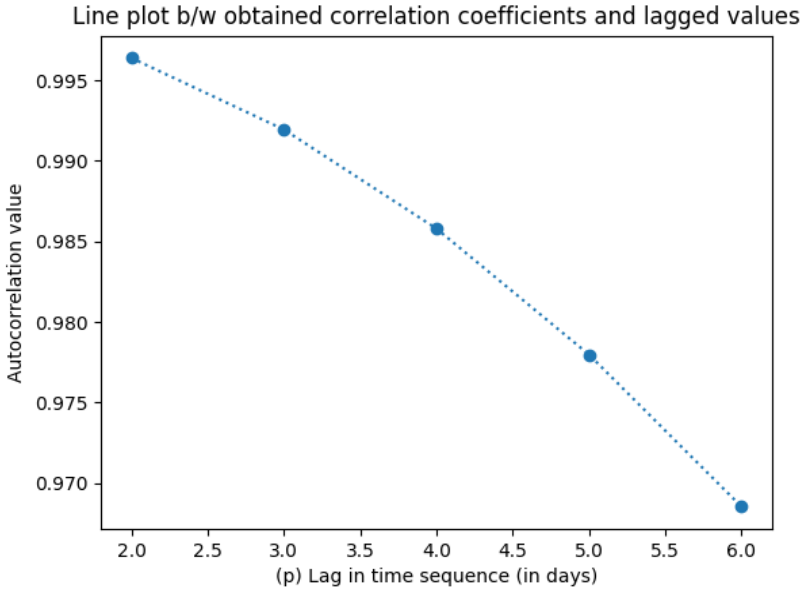

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation coefficient values seem to be decreasing when the value of lag for the time-sequence is increased.
2. The reason for the observed declining trend in correlation value when we increase the lag in time-sequence is because the trends in the given data (covid- cases) are quite gradual not sudden or abrupt. Thus, the value at any given point is expected to be dependent on its immediate previous value more than any value much further away from it.

e.

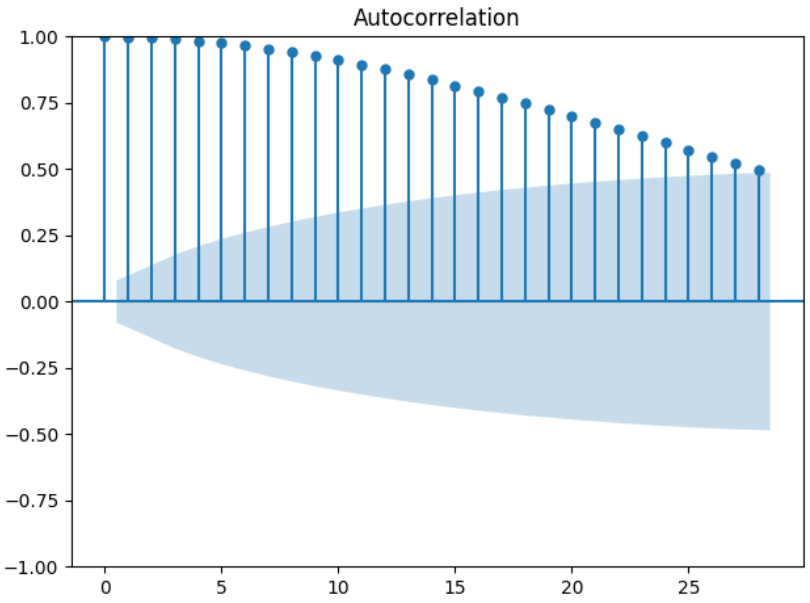

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. It may be observed that the autocorrelation value is decreasing when we increase the lag in the given time series data.
2. The reason for the observed declining trend in correlation value when we increase the lag in time-sequence is because the trends in the given data (covid- cases) are quite gradual not sudden or abrupt. Thus, the value at any given point is expected to be dependent on its immediate previous value more than any value much further away from it.
3. Thus, if the lag is further increased the autocorrelation value would further decrease as the data would become more and more dissimilar to the present data.

2

a. The coefficients obtained from the AR model are;

- ✓ The value of w0: 59.955
- ✓ The value of w1: 1.037
- ✓ The value of w2: 0.262
- ✓ The value of w3: 0.027
- ✓ The value of w4: -0.175

✓ The value of w5: -0.152

b. i.

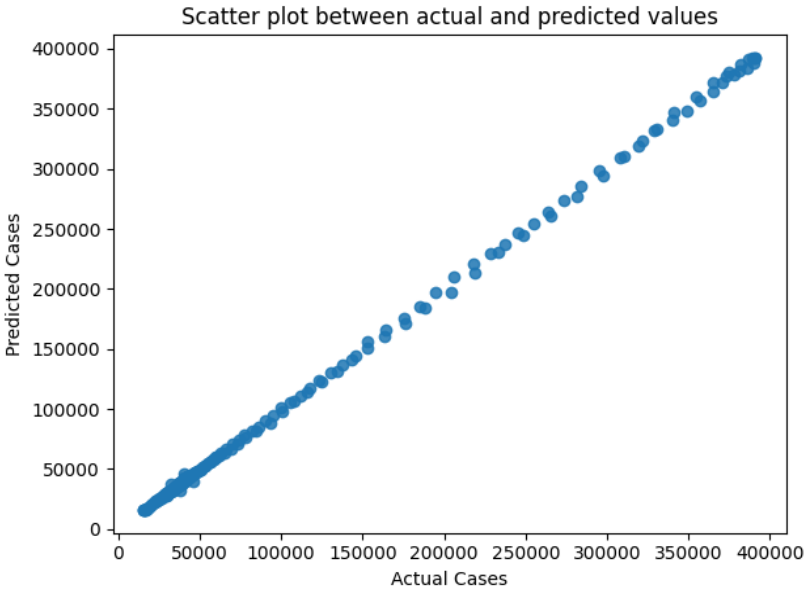

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From observing the scatter plot, we may infer that the two-time sequences are strongly positively correlated with each other.
2. The scatter plot exactly follows the nature reflected by Pearson's coefficient in the 1.b.
3. It may be clearly verified that the present data has some sort of dependency on the previous data values. Thus, when the previous values are taken as inputs, we may observe that the predicted and the actual values shows quite a good positive correlation thus in conclusion we may say that our assumption is correct that there is dependency on previous values in the given data.

ii.

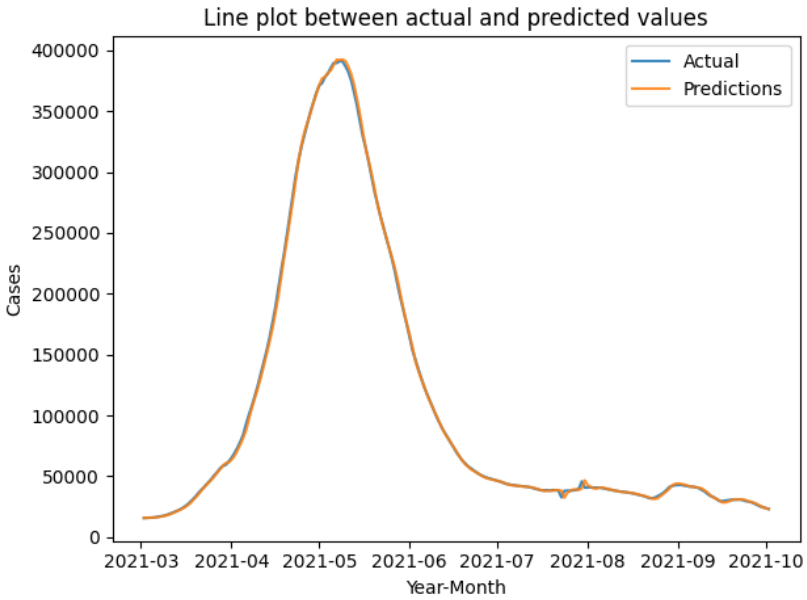

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. We may observe that the actual and predicted cases are nearly same over the given time-period.
2. There is a high degree of positive correlation observed between the actual values and the predicted values. This shows that our predicts are showing a high dependency on the actual data present as a very similar trend is observed for both. (Both seems to be increasing at nearly same pace).
3. Thus, this proved that our model predictions are very much accurate and thus this model can be used for making any further future prediction.

iii. The observed vales of RMSE (%) and MAPE are:

RMSE (%) :1.827

MAPE: 1.589 %

Inferences:

1. The error is very small and thus the accuracy can be said to be very high and the model can be assumed to be very much reliable for future predictions.
2. As the present data is showing a high dependency on the previous on the previous values, hence it was to be observed that our model will make very good predictions when given inputs with previous lagged values. That's why a high accuracy and a low error is observed.

3

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

Lag value	RMSE (%)	MAPE
1	5.379	3.486
5	1.827	1.589
10	1.688	1.533
15	1.614	1.51
25	1.706	1.552

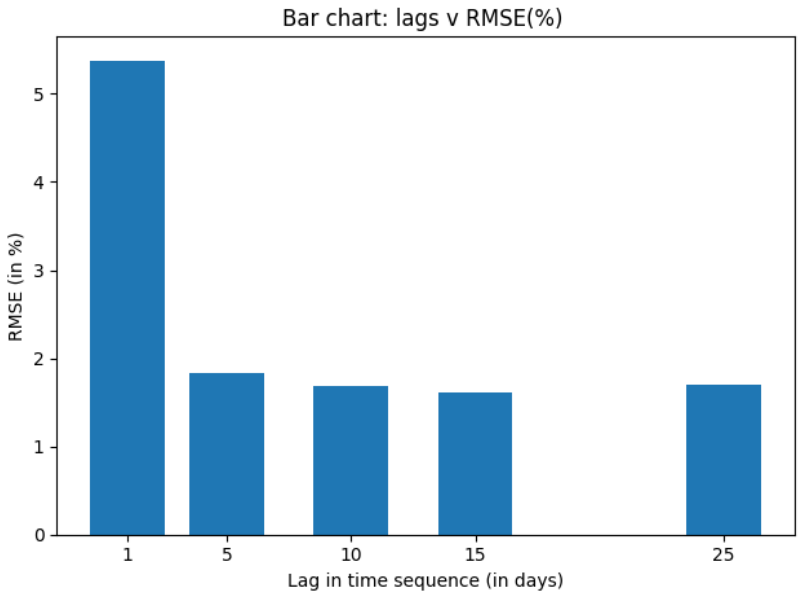

Figure 7 RMSE(%) vs. time lag

Inferences:

1. With increasing lag values, the RMSE (%) values are decreasing at first, and then the change becomes pretty much saturated. A sudden decrease is observed in the beginning though.
2. This is observed because the target values show the dependency to some acceptable correlation value up to some lags only. After that the dissimilarity starts to increase and thus our model is not able to increase the accuracy further.

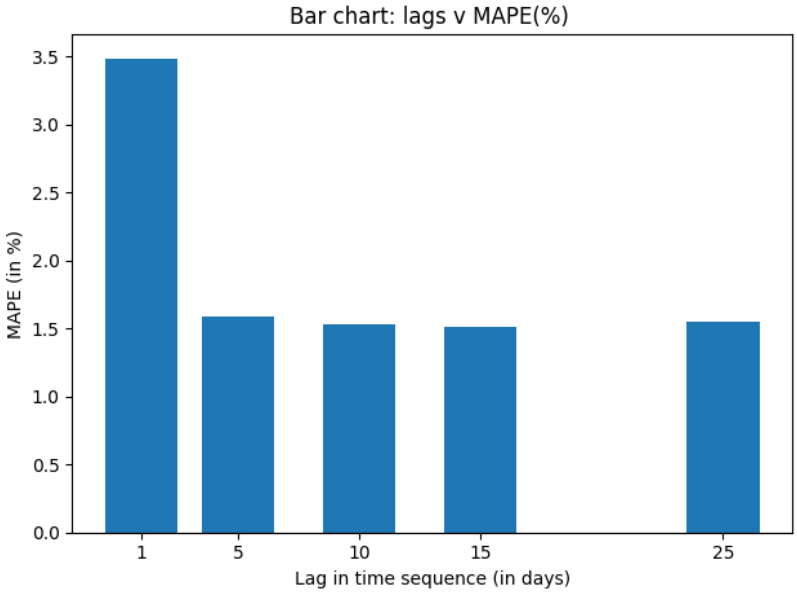

Figure 8 MAPE vs. time lag

Inferences:

1. A very similar or we may say same trend is observed here as observed with RMSE (%) values previously.
2. With increasing lag values, the MAPE (%) values are decreasing at first, and then the change becomes pretty much saturated. A sudden decrease is observed in the beginning though.
3. This is observed because the target values show the dependency to some acceptable correlation value, up to some lags only. After that the dissimilarity starts to increase and thus our model is not able to increase the accuracy further.
4. The lag values 15 and 25 days have nearly same MAPE (%) values.

4

The heuristic value for the optimal number of lags is 77 days.

The RMSE (%) and MAPE (%) value between test data time sequence and original test data sequence are 1.763 and 2.068.

Inferences:

1. The heuristics for calculating the optimal number of lags have slightly increased the best values obtained for lag of 15 days in the previous problem. The difference is though very small and can be easily neglected. We may say that the heuristics have provided a good result, but the same or slightly better result could have been obtained out of much smaller lag of 15 days. This also computationally less intensive.
2. The reason for this observation is that as we increase the lag values in a given time-series data, the dissimilarity with the target data point (to be predicted) increases. As the process of change in values in such sequences is gradual not sudden/abrupt. Thus, only up to some optimum number of lags which have an acceptable correlation with original sequence produces satisfying and realistic results. Taking too much lag would be resulting in including inputs that show zero or no correlation with the original data and thus making false or unreliable predictions. Thus, lag value needs to be chosen wisely.
3. Previously the optimal value of lag was 15 days having the least possible errors i.e., RMSE% was 1.614, MAPE was 1.51 %.
4. On comparing we may see that the lag values calculated without heuristics gives slightly better results than the one calculated with heuristics. It is because a lot of lag values have been taken in heuristics method that have a very low or zero correlation with the original data and hence it has somewhat less accuracy.
5. But the gap between the results is too small that heuristics model can not be neglected easily.