No. of group members - 5

Sanket Takle

Priyanka Biswas

Krina Parmar

Gayatri Patil

Ashish Ingale

Problem Statement

Perform Time Series analysis and forecasting in order to study the impact and spread of the COVID-19 in coming days with the help of number of confirmed cases based on past data.

Data collection and Processing

We searched for the dataset on several sites for trending topics. We found out one dataset on COVID-19 containing around 10000 observation and 9 features. It contained Date feature which is number of days from 22nd January 2020 to 5th June 2020, confirmed cases, death cases and recovered Cases which were important to us

In data preprocessing, first we made the dataframe containing only date column and confirmed cases. We plotted the line chart to see the

frequency of our data. We had cumulative data in our dataset so we had to convert it into daily cases for better forecasting.

After that we split the dataset into train set and test set. We checked for stationarity in our data with the Augmented Dickey Fuller test in which we test the T statistic with the critical values for rejecting the null hypothesis

We performed decomposition method to check the trend and seasonality in our data. We plotted the tsa. ACF and PACF plot to check the auto correlation function and partial auto correlation function which helped us to select the seasonal order in our model.

We performed modelling on SARIMAX which is AR specification, Integration order, MA specification and AUTO ARIMA. In this problem and data set SARIMAX performed well with mean absolute value of 89.07