Statistical Machine learning, Forecasting and Inferences from Pandemic Data (India)

Project Report-2020

A project report submitted in partial fulfillment of the requirements for the Project in Data Analytics

Integrated Post Graduate (2017-2022)

by

D.Sravan Kumar (2017IMT-030)

Submitted to-

Dr. Anuradha Singh

ABV INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR - 474015

2020

TABLE OF CONTENTS

1. ORIGIN OF THE PROPOSAL

Coronavirus 2019 is powerful contamination that began in China in late 2019, spread swiftly around many countries in the world. It has had a critical effect on the public's way of life and also on the world economy. The financial and social disturbance caused by the virus is annihilating: millions of individuals are at the hazard of falling below the poverty line. The Covid-19 has influenced communities, businesses, and associations all-inclusive, incidentally influencing the money related markets and the global economy. The Rise-up of COVID-19 has made a calamitous circumstance all through the world.

The average rate of Coronavirus is exponentially increasing day-by-day. So,there is a growing need to manage and analyze big data of infected persons, patient details, their community gatherings, and integrate them with clinical trials and public health data. Machine Learning models can be used exceptionally well to successfully track the virus, and can also be used to infer future development of the virus, and accordingly plan procedures and approaches to prevent its spread. This data also helps administrative units across the world to implement necessary precautions to prevent the spread. By and by there's a chance of a second wave within the nations just like the U.S.A, India, Brazil, etc. There's a vital requirement of foreseeing long haul circumstances and taking essential measures in checking the misfortune of human life as well as financial effect. We propose an approach in anticipating the long-term circumstance of India in terms of no of cases, Death rate, Recovery Rate, utilizing measurable demonstrating with time arrangement determining examination.

2. REVIEW OF STATUS OF RESEARCH AND DEVELOPMENT IN THE SUBJECT

2.1 International Status:

The severe problems posed by several pandemics in the past have led statisticians and researchers around the world to take up the current coronavirus pandemic very seriously and maintain data related to it to analyse the situation and predict future possibilities in terms of number of people affected, number of deaths, etc. so that governments can take appropriate action to curb the spread of the coronavirus pandemic to the maximum extent possible while at the same time preserve their respective economies and ensure smooth functioning of other activities.

Various models have been developed to predict future data regarding the above mentioned based on existing data of the same. For example, in the Hubei Province in China, a network model which goes by the name of Network-Inference-Based Prediction Algorithm (NIPA) is used to predict the future information regarding the pandemic wherein the network comprises of the different cities in the province and the interactions between these cities like traffic flow, etc. are then used to make these predictions. Elsewhere in the country of South Africa, a model named Susceptible Exposed Infected Recovered (SEIR) and Susceptible Infected Recovered (SIR) epidemiological models with time dependent trending rates based on Bayesian parameter estimate using Markov Chain Monte Carlo (MCMC) methods which has been used to forecast the future data. All these mathematical models help in taking preventive measures to stop the rapid spread of the virus in future based on the predictions that are obtained as a result of these models.

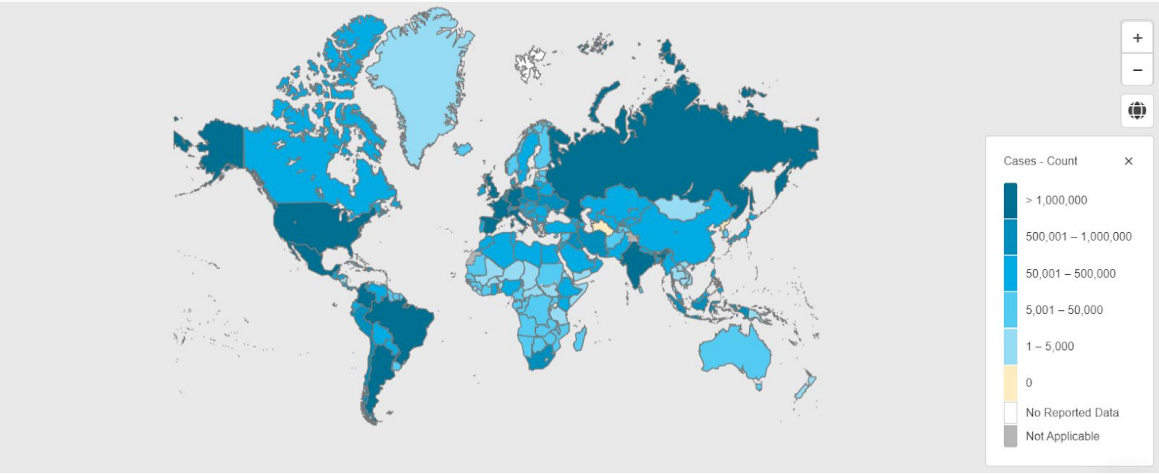


Figure 2.1: Worldwide classes of infections (Source: WHO)

The advantage of maintaining this data, developing these models and sharing those results is that it helps in controlling the spread of virus in countries where the situation is in the earlier stages by making use of the data from countries where the situation has gone a bit out of hand. For example, when the effect of the virus was still in its primitive stages in India, the data collected from other countries like Italy, etc. helped in taking measures like imposing a lockdown across the country which proved to be quite an effective measure in curbing the rapid spread of the virus. Also, based on the predictions made, there is expected to be a second wave of the pandemic which has led certain countries like London to impose a second phase of lockdown in their country which is in effect currently. These predictions also help governments to plan their next moves and take the appropriate steps keeping the country's economy and other such important factors in mind.

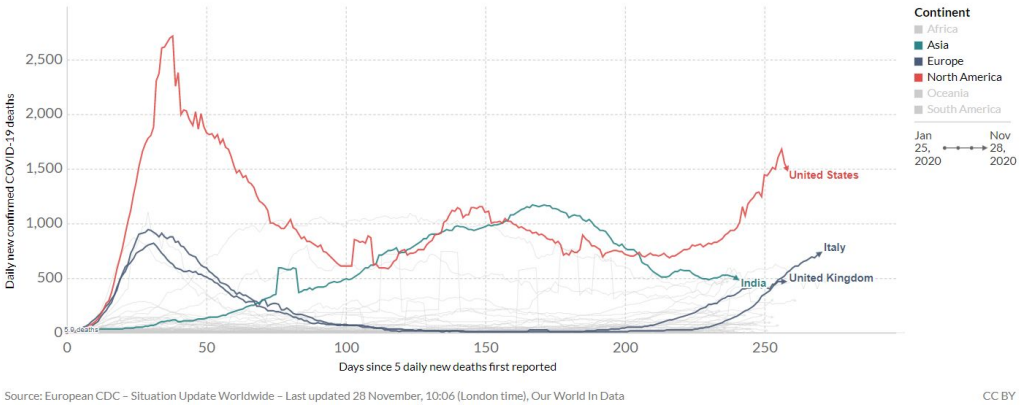


Figure 2.2: Infection rates of worstly affected countries (source: ourworldindata)

As of 23rd November 2020, only two vaccines namely Sputnik V (Earlier named as Gam-COVID-Vac) and EpiVacCorona have been approved and both of these have received approval from the department of Ministry of Health in the Russian country. These vaccines have however raised several concerns regarding safety and efficacy as both these vaccines have not entered phase 3 trials. Several US federal government departments have collaborated to form Operation Warp Speed(OWS) and have selected three vaccines to invest for phase three trials namely University of Oxford and AstraZeneca's AZD1222, Pfizer and BioNTech's BNT162 and Moderna's mRNA-1273.

2.2 National Status:

Various analyses on the spread of covid-19 in India were conducted initially by several researchers using the available official data. The generic MCMC methods were not applicable to model india data at the time of April 2020, Arithmetic models were explored which involved using a “change-factor” or “rate-of-change” algorithms, the models used in other countries were helpful for modelling india data at the time of initial stage in April 2020 during lockdown period, hence for india early related works were mostly based on mathematical approaches with variations amongst them. The ratio of the daily increase with respect to the past *N* days is estimated and used for developing the model and making forecasts by a team of researchers. The lower and upper bounds are also provided by such models as the actual prediction might be hard to estimate in such a pandemic situation as any presence of cluster or community transmission will spike up the cases.

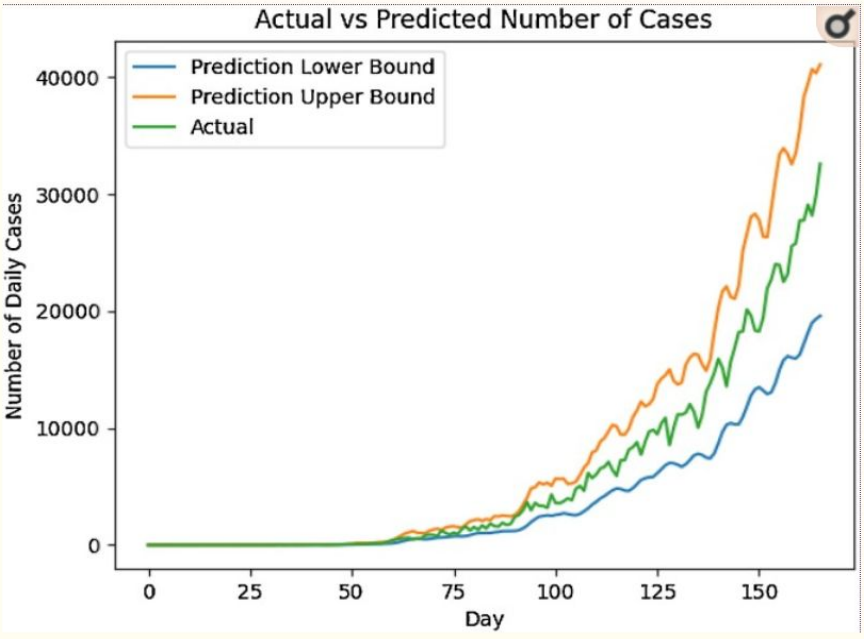


Figure 2.3: Forecast made by the one of existing model for india (Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7445130/)

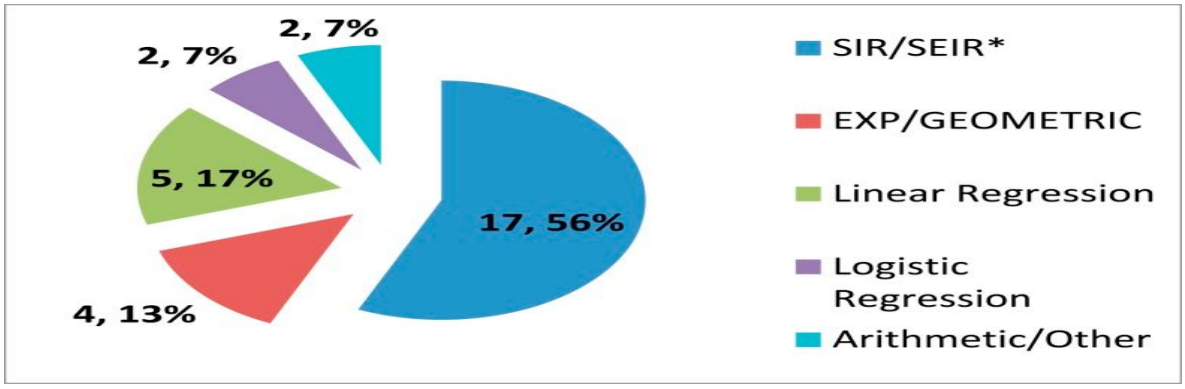


Figure 2.4: Distribution of models used for modelling covid spread in india
(Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7298493/)

Among all the various math models used in tracking the spread in India the distribution is given in the above figure, the latest data used for mathematical modeling was as of 13 April 2020. The assumptions were made on pandemic modelling like estimation of R0 and other parameters, using python and R programming languages and visualizing the results and using it for making government decisions were initially done in the context of national status for statistical machine learning and inference from pandemic data. The data used were from distributed sources like WHO, worldometer, John-Hopkins university etc.. Majority of models were based on Susceptible Infected Recovered (SIR).

2.3 Importance of Proposed Project :

This project focuses on the analysis and drawing inferences using statistical machine learning of pandemic data for india.

Future Analysis of Covid-19 is really important in preventing the loss of human lives and estimating the economic impact of different countries across the globe. Many countries have been able to tackle Covid-19 by imposing lockdown ahead than other countries. Countries like New Zealand, Korea, Thailand, etc. have saved themselves by imposing lockdown and taking necessary measures by analysing the Statistical models, time series forecasting models and taking necessary preventive steps and medical facilities establishment.

With a decrease in the number of infected cases, many countries across the world removed their lockdown restrictions. However, the recent resurgence of COVID-19 cases in Europe has again forced nations to retrace their steps back to imposing lockdown measures to control the spread of covid-19.Therefore, the second wave of a Covid usually indicates to a previous Covid infection that seems to be decreased for a certain period of time and increases after a certain period, resulting in a second wave of infections, deaths and huge economic loss across the globe.

In India, the new infected cases and number of deaths have decreased periodically over the last three weeks and the pandemic situation has been somewhat stabilised in most of the states, There may be a one more chance of a second wave that looms over the world. Therefore, the analysis of the current situation will help flatten the curve by taking necessary steps and aligning the government policies to the same, and also be prepared in case a second wave of infections is predicted to be the outcome of the analysis.

Hence, We propose an approach using the SARIMAX statistical time series model in analysing the future situation of India in terms of no of cases, no of deaths and no of Recovery cases in near future based on covid-19 data.

3. WORK PLAN

3.1 Methodology

Dataset Description:

The dataset is gathered from a crowd sourcing initiative by Covid19India organisation team. The dataset contains details of covid cases with timestamps on a daily basis. The sourcing of the data is from various news articles, media, social media handles of official state governments. Availability of such open source data helps largely for various analysis purposes to the governments and researchers to get insights about the situation. Though the data is not official it is largely used by many indians and is accurate up to date.

The dataset is collected in CSV format with information of number of active, recovered cases along with number of testing and death information statewise and for every day in a separate row. The information is contained from the start of the epidemic in India as of 1 case on January 30th, 2020 to 92,25,045 cases as of Nov 24th. We have utilised the API endpoints built by the team to gather the data and draw our analysis and insights from it.

Dataset Acknowledgment:

- https://www.covid19india.org/ (COVID-19 India Organisation Data Operations Group)
- https://api.covid19india.org/
- https://www.mohfw.gov.in/ (Indian Govt. Ministry Official website for covid data)
- https://www.cdc.gov/coronavirus/2019-nCoV/index.html (CDC of the US Govt.)

SARIMAX model:

Seasonal Autoregressive Integrated Moving Average, SARIMA or seasonal ARIMA, is an extension of ARIMA that supports univariate statistical information with a seasonal element. Typically, Arima is employed to research the statistical information however with a further element of seasonality in SARIMAX , we've chosen the SARIMAX model to research our statistical information that has seasonality in it.

It has 3 new hyperparameters known as the autoregression (AR), differencing (I) and moving average (MA) for the seasonal element of the series, it also has an additional parameter which is called seasonality.

A seasonal ARIMA model is made by as well as extra seasonal terms within the ARIMA The seasonal part of the model consists of terms that are terribly the same as the non-seasonal parts of the model, however they involve backshifts of the seasonal amount.

Seasonal Elements:

There are four seasonal components that are part of SARIMA model are:
P: seasonal autoregressive order.
D: seasonal difference order.
Q: seasonal moving average order.
m: the quantity of your time steps for one seasonal period.

Together, the notation of the SARIMA model is defined as: Sarima(P,Q,D)m. Significantly, the m parameter is a key component for the P, D, and Q parameters. M indicates the period of the seasonal cycle. The trend components may be chosen through an effective study of ACF and PACF plots viewing the correlations of recent time steps . Similarly, ACF and PACF plots may be analyzed to specify values for the seasonal model by viewing correlation at seasonal lag time steps.

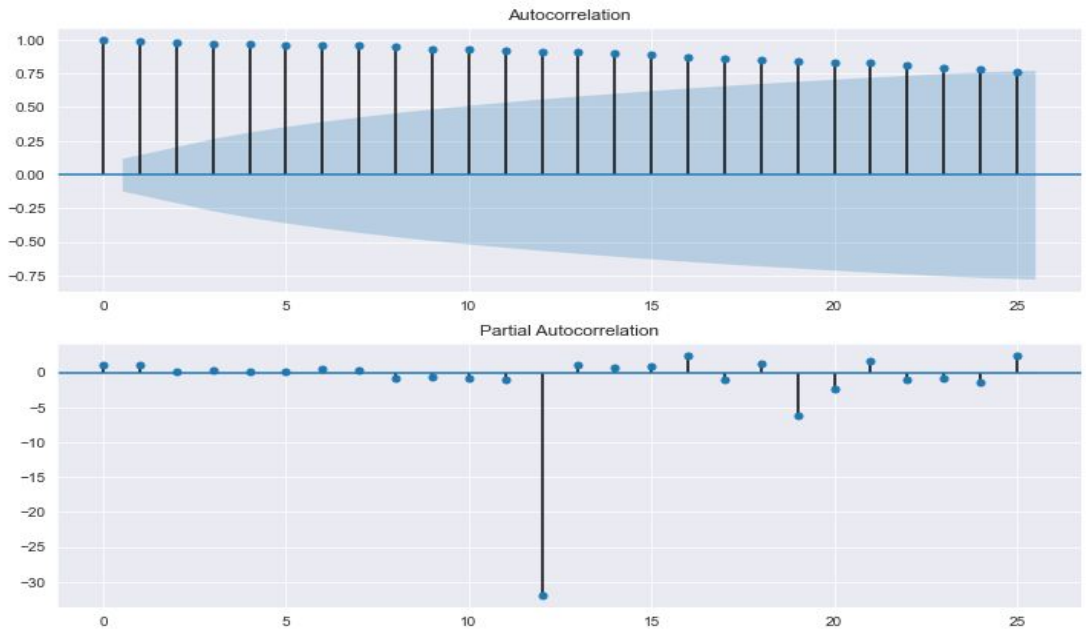


Figure 3.1: Correlation (ACF, PACF) Analysis of covid-19 india data.

Here in partial autocorrelation, we can see a small upward trend for (almost) every second plot as compared to the previous plot. Therefore our data is not stationary inorder to test stationarity of our data we have used the Augmented Dickey-Fuller unit root test.

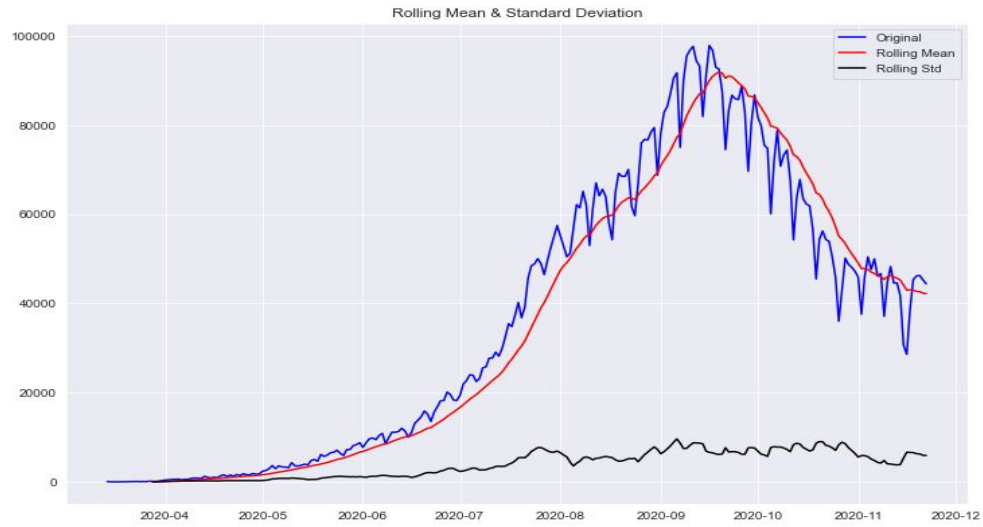
Dickey Fuller Test:

Statistical tests build robust assumptions concerning your information. all the same, they'll give a fast check and confirmative proof that some time series is stationary or non-stationary. The increased Dickey-Fuller test could be a form of applied mathematics test known as a unit root test. The theory behind a unit root test is that it describes how powerfully a statistic is outlined by a trend. The Dickey-Fuller test employs an autoregressive model and optimises an information criterion for multiple different lag values.

The null hypothesis of the test is that a unit root should represent the time series, that it is not stationary (has some time-dependent structure on the data). The alternative hypothesis (rejection of the null hypothesis) is that the time series is stationary in data.

- Null Hypothesis (H0): If it is not dismissed, it shows that the time series has a root unit, which implies that it is non-stationary. It has a structure based on a certain time.
- Alternate Hypothesis (H1): The null hypothesis is rejected; it implies that there is no unit root in the time sequence, which means that it is stationary. It does not have a function depending on time.

We have used Dickey Fuller Test to test our stationarity in our data. We have used a level of significance alpha-0.005 in performing this test on our data in order to evaluate our Null and Alternate Hypothesis.




```
Results of Dickey-Fuller Test:
p-value = 0.4484. The series is likely non-stationary.
Test Statistic                  -1.666513
p-value                          0.448443
#Lags Used                      15.000000
Number of Observations Used    238.000000
Critical Value (1%)             -3.458128
Critical Value (5%)             -2.873762
Critical Value (10%)            -2.573283
dtype: float64
```

Figure 3.2: Dickey Fuller Test for Stationarity check indata and its level of significance(p)

Here we can observe the huge gap between original data and Rolling mean, Rolling standard deviation in the figure 3.2 and it also states the p value is 0.448443 which is not so good and hence , the output says "The series is likely non-stationary".

There are various methods for making series stationary like log, differencing and so on, here we are using Differencing , shift operator shifts the independent column of dataframe by 4 places and difference is taken.


```
Results of Dickey-Fuller Test:
p-value = 0.1093. The series is likely non-stationary.
Test Statistic                  -2.525797
p-value                          0.109323
#Lags Used                      15.000000
Number of Observations Used    234.000000
Critical Value (1%)             -3.458608
Critical Value (5%)             -2.873972
Critical Value (10%)            -2.573396
dtype: float64
```

Figure 3.3: Dickey Fuller Test after performing differencing Technique to stationarize the data and its level of significance(p).

Plotting the data after differencing we see the p value is reduced to 0.109323 which is quite good as compared to our previous value 0.448443. We can try different values in shifts to reduce the p value , we have used only 4 places to shift in order to stationarize.

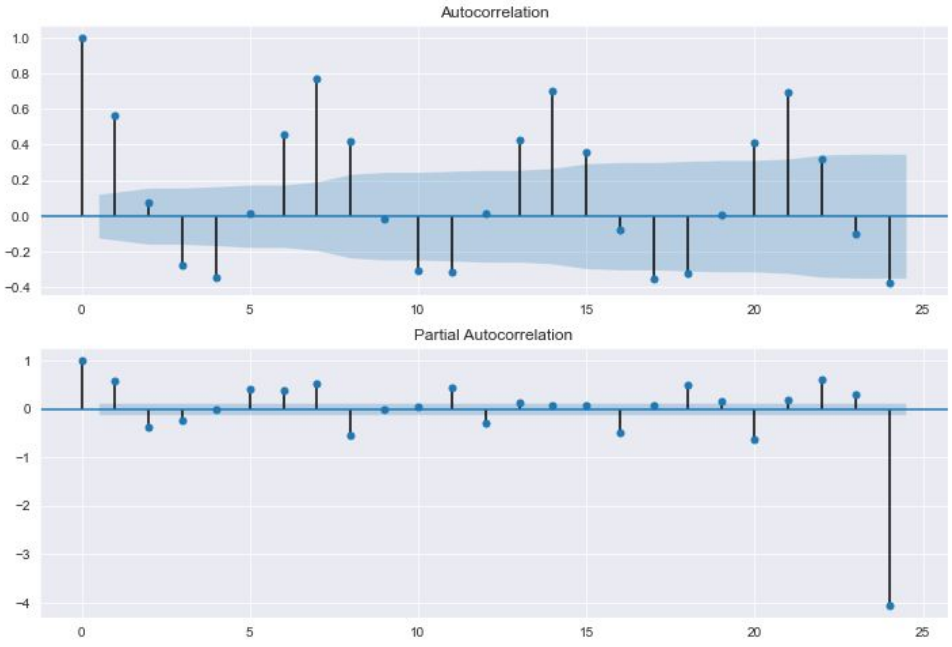


Figure 3.4: Correlation (ACF, PACF) Analysis after stationaring data using Differencing.

Now we can see the autocorrelation and partial correlation of our data. After reducing the p value to 0.109323 we can see the upward trend of our data has decreased significantly and seasonality has improved periodically.

We can see a recurring correlation exists in both ACF and PACF hece we should choose the SARIMAX model which also deals with seasonality.
A model with no orders of differencing assumes that the original series is stationary (mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend. A model with two orders of total differencing assumes that the original series has a time-varying trend.
Since our series has a constant average trend (with growth) we would take D = 1 and Q= 0 (D-1) and we have chosen P=14.

SARIMAX Results

Dep. Variable:	TT	No. Observations:	254
Model:	SARIMAX(14, 1, 0)	Log Likelihood	-2314.419
Date:	Wed, 25 Nov 2020	AIC	4658.837
Time:	18:38:12	BIC	4711.838
Sample:	03-14-2020 - 11-22-2020	HQIC	4680.161
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2329	0.049	-4.750	0.000	-0.329	-0.137
ar.L2	-0.0888	0.061	-1.454	0.146	-0.208	0.031
ar.L3	0.0599	0.060	0.995	0.320	-0.058	0.178
ar.L4	-0.0713	0.053	-1.339	0.181	-0.176	0.033
ar.L5	-0.0115	0.054	-0.213	0.831	-0.117	0.094
ar.L6	0.1167	0.066	1.772	0.076	-0.012	0.246
ar.L7	0.5969	0.055	10.924	0.000	0.490	0.704
ar.L8	0.1861	0.058	3.224	0.001	0.073	0.299
ar.L9	-0.0138	0.069	-0.202	0.840	-0.148	0.121
ar.L10	-0.0836	0.063	-1.330	0.183	-0.207	0.040
ar.L11	0.0096	0.067	0.144	0.885	-0.121	0.140
ar.L12	-0.0478	0.051	-0.932	0.351	-0.148	0.053
ar.L13	-0.1123	0.062	-1.813	0.070	-0.234	0.009
ar.L14	0.2357	0.057	4.141	0.000	0.124	0.347
sigma2	5.283e+06	3.09e+05	17.102	0.000	4.68e+06	5.89e+06

Ljung-Box (Q):	63.80	Jarque-Bera (JB):	160.76
Prob(Q):	0.01	Prob(JB):	0.00
Heteroskedasticity (H):	47.74	Skew:	-0.45
Prob(H) (two-sided):	0.00	Kurtosis:	6.80

Figure 3.5: SARIMAX model training results and parameters used.

Therefore we have performed SARIMAX time series model on our data with seasonal elements P,D,Q,m. We have chosen SARIMAX(14,1,0) as P,D,Q values respectively.

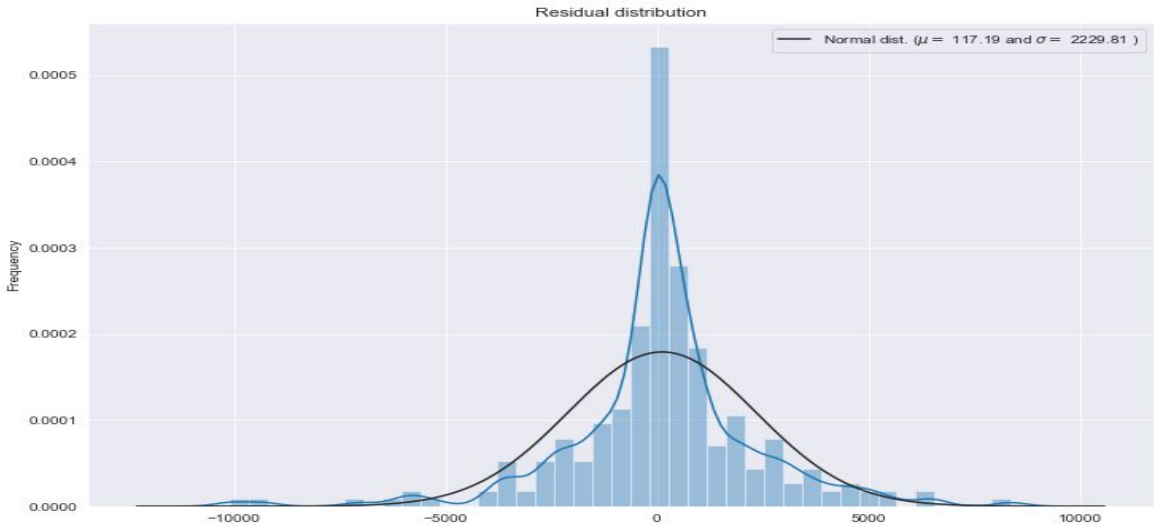


Figure 3.6: Data distribution plot after applying residuality to our SARIMAX model

We have used residual distribution to our model which converts our model to normal distribution with U and sigma where u=, sigma= . in order to decrease the remaining upward trend in the data and making our data more stable. We can also see the autocorrelation and partial correlation in the plots which gives us a more robust trend.

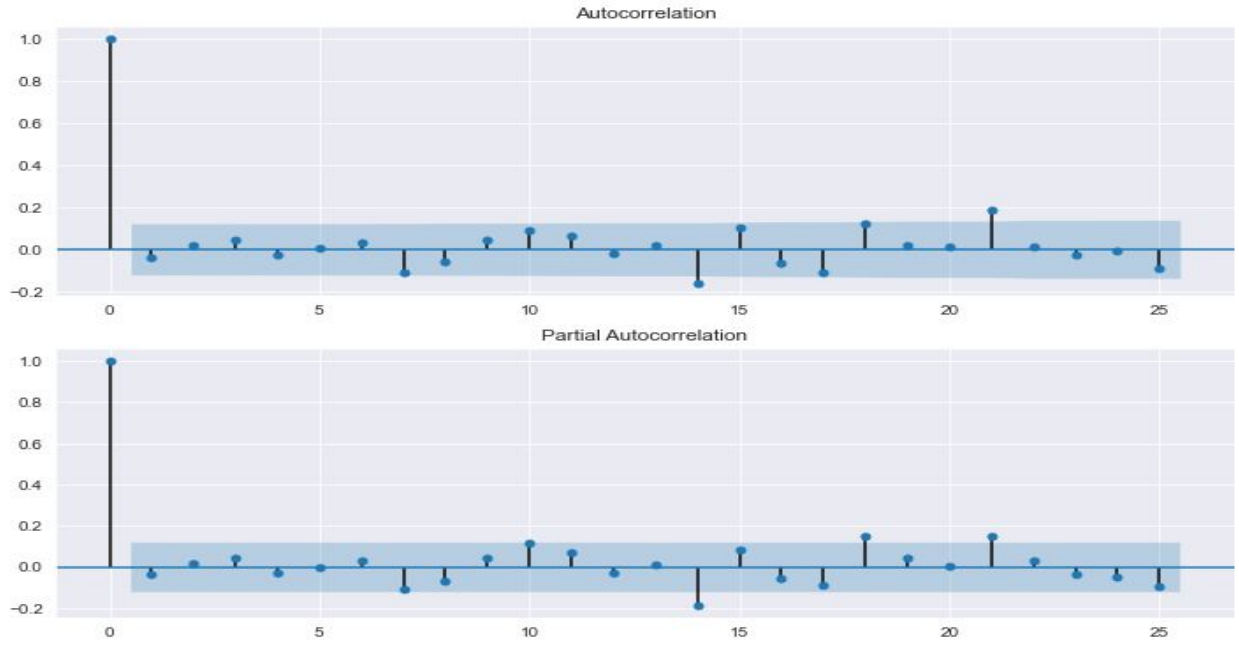


Figure 3.7: Correlation (ACF, PACF) Analysis after applying residuality to our SARIMAX model.

Finally we have successfully builded our model using the design principles which have been discussed above. We have built our mathematical time series forecasting model using the SARIMA time series model with necessary parameters.

3.2 Results

The model is constructed such that it is able to fit a possible curve that will predict the situation in future of India, on the quantities of new infection population, recovery rate and an inference is drawn based on the model's estimation on weather a 2nd wave is bound to happen in India as in other countries, or will it be possible to flatten the curve without such situation.

The time series forecast advantage is that it is able to do realtime forecast as the data is being fed into the model and computations made by SARIMAX model have been widely used in such pandemic and epidemic modelling.

The plot of the predicted infection rates by our model upto date is shown as below:

Figure 3.8: Forecast upon the actual data upto date from April 1, 2020

This forecast is made by the learned model based on the data upto each point in time, it is pretty close to the actual value and using this kind of model to make future insights would be efficient and accurate upto accepted level. The model is able to imitate the curve closely which is a means that we are in possible correct direction for making a forecast.

The forecast by our model for the next 30 days of Infection cases in India is shown below:

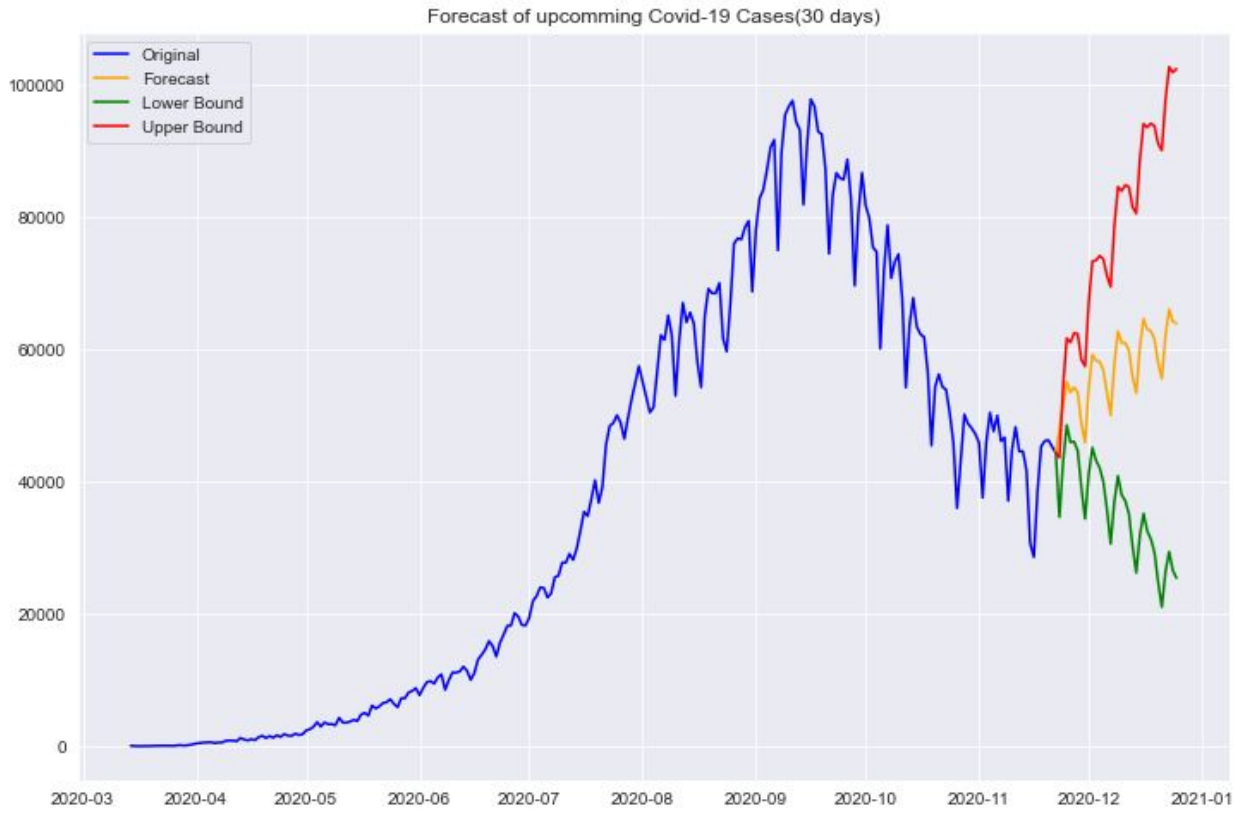


Figure 3.9: Forecast of upcoming Covid-19 Cases in india (30 days)

The forecast for December 2020-January 2020 indicates that the new infected population would be increasing significantly, the upper and lower bounds are also estimated, if the new cases follow the pattern close to the lower bound curve, it is good to assume that the cases are being significantly reduced despite the seasonal situation in India. The lower and upper bound values are estimated based on the confidence interval chosen by the model while training which fits best.

The forecast for the recovery rates are also determined by our model, the plot of estimated recovery rate upon the actual is shown below:

Figure 3.10: Forecast of recovery rate upon actual data

The forecast on the actual data is closer and the model is doing good for recovery rates too, and hence it could be used to foresee the next 30 days of recovery rate in India.

The forecasted recovery cases for December 2020 - January 2021 by our model:

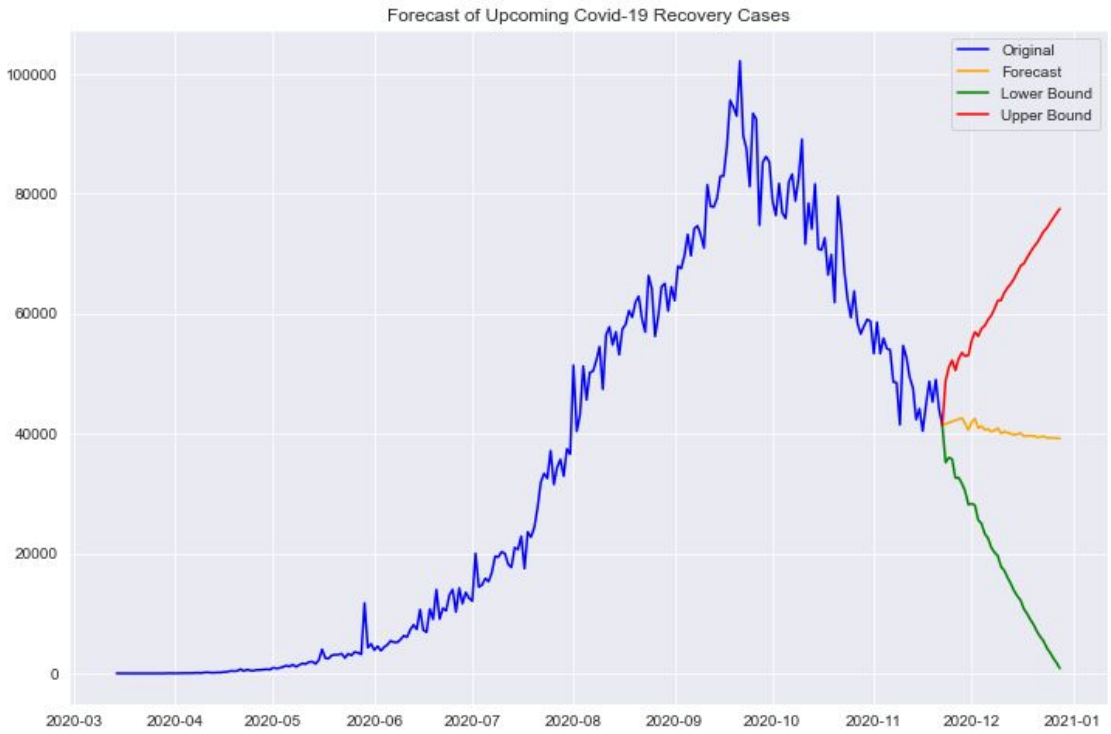


Figure 3.11: Forecast of upcoming Covid-19 Recovery Cases (30 days)

The recovery rate for the next 30 days are estimated to be almost flat and no much observable change in the rate, however the lower and upper bound of the curve is an indication that if the pattern follows red trend the number or recovery rate could surpass the new infection rates paving a way to flatten the curve as hoped.

The forecasted deaths for December 2020 - January 2021 by our model:

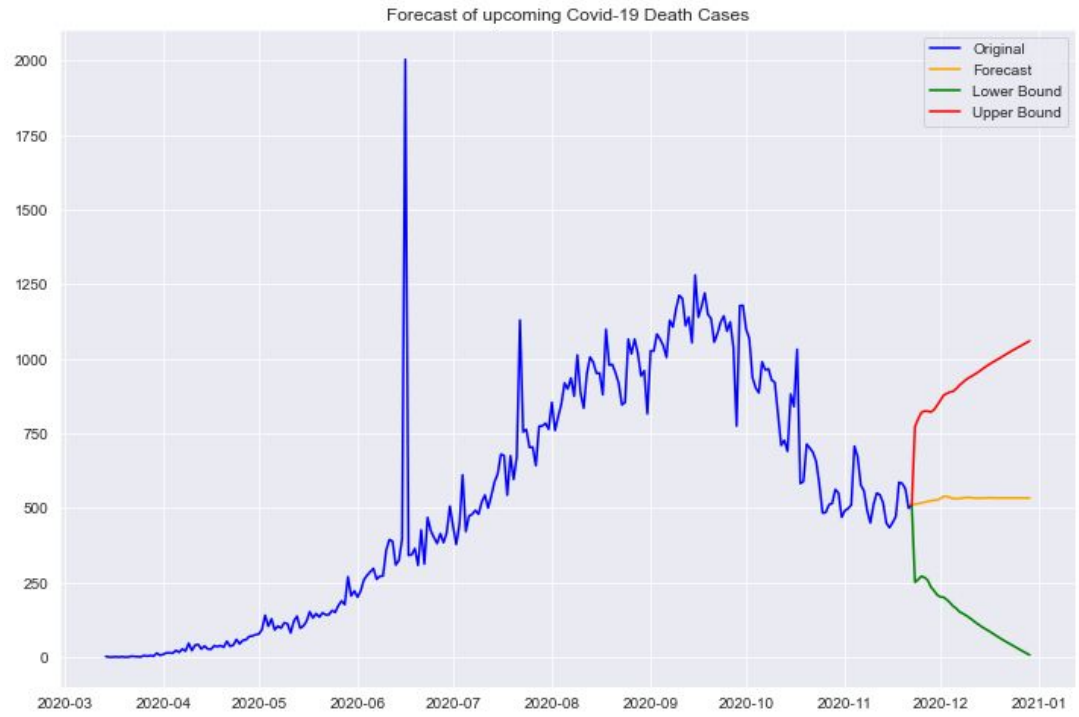


Figure 3.12: Forecast of upcoming Covid-19 Death Cases (30 days)

3.3 Conclusion

Our study on the covid-19 data in india based on statistical machine learning, forecasting and inferences which showed how important it is to analyse the data available and get prepared ourselves for the next situation. The results shown for the 30 day period is an indication of what direction we could be heading into, though it is not completely accurate as the actual estimation will be hard due to many unknown factors prevailing in the pandemic situations. Our time series model which used SARIMAX have predicted that there is strong chance of second wave going to come in near future, but this second wave has less impact on our lives compared to the first wave as our model is showing only decent amount of increase in cases which may be 60000 cases per day approximately in India. Based on this analysis we want to conclude that prevention is always better than cure, so taking necessary conditions by the government and ourselves is very much important in the near future to save ourselves from this pandemic situation. Many countries have been successful in restricting the coronavirus by taking necessary measures. From this statistical inference India should be very cautious in handling coronavirus and get prepared for the future situation in terms of equipment, Hospitals or any other essential resources including manpower and improve the rate of speed of vaccines invention is very much important to save from this deadly virus.