Modeling a Pandemic : Effectiveness of Different Models

Nomula Rohith Yogi Rohith.Yogi@iiitb.org

Bukka Nikhil Sai Nikhil.Sai@iiitb.org

Dachepalli Soumith Kumar Soumith.Kumar@iiitb.org

Abstract

COVID-19 (SARS-Cov 2) is currently a global concern. With raising number of cases, deaths and economic impacts due to lock downs, it is fair to say that COVID-19 is the major problem facing mankind today. India is a densely populated. In this stage, modeling the spread is very important and crucial for decision making. In this paper, we test the most used epidemiological models (SIR, SIER) and compare them with a normal forecasting model (Prophet) too see how effective these models are in predicting the spread of Covid-19 in India. We also include various analysis performed on publicly available data that could be used to improve the performance of the models.

1 Introduction

With new cases of COVID-19 being reported everyday in India, it is very important to analyze the spread to gain insights on how this could be prevented. Indian government has made the COVID-19 data publicly available (MoHFW) and there are other public crowd sources data sources. These resources can be utilized to obtain demographics of current Covid outbreak and to understand exactly, how we got here. These insights can be used to model future spread and to see how the measures that are currently being implemented like social distancing, movement restrictions and lockdowns are working. The objectives include exploratory analysis of the available data to study the phenomenon, to draw patterns enabling us to model it. Few of the models we used are presented in this paper.

Epidemiological modeling of the spread is crucial for decision making but the question remains about the effectiveness of these models. Covid outbreak gives us a chance to test this. In this paper, we present a few famous epidemiological and forecasting models like SIR [1], SIER and how effective they were in modeling Covid spread. For this purpose, we try to model the spread in India over a period of time using the publicly available data to gain some insights into the performance of these models.

Our contribution is comprised of but not limited to the following

- Study how the virus has spread in the initial phase.
- Model the growth trend, understand the factors influencing it.
- Use mathematical models for forecast.
- To understand the impact decisions based on this model could make.
- Compare the effectiveness of various models. [2].

The rest of the paper is structured as follows. In the next section, we present various analytics we have performed to get insights from the data to set various parameters in the models we have used. This is followed by a discussion about the various models we have used along with the results.

2 Data Analysis

Our data set is collected from official COVID-19 India web page where the patients are tracked state wise.[3] This data is crowd sourced and is available online for every one via api and can be downloaded in JSON or CSV format. From this available data sources we have taken into account the complete patients details, state wise testing details and also population of India according to 2020. After pre-processing the data we have grouped different analysis on the data into different sub sections.[4]

2.1 Spread Analysis

From 1 and 2 we can see that the Total Number of Confirmed cases has reached a maximum of around 95k during the lockdown (March 2020 to May 2020) and we were in a better position when compared with the other countries.As soon as the lock down relaxation started the number of cases increased significantly more when compared to other countries and reached around 600K on 30th June 2020.

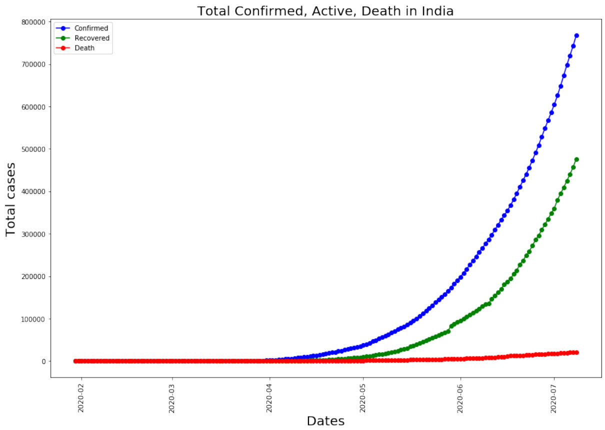

Figure 1: Covid-19 Cases Distribution

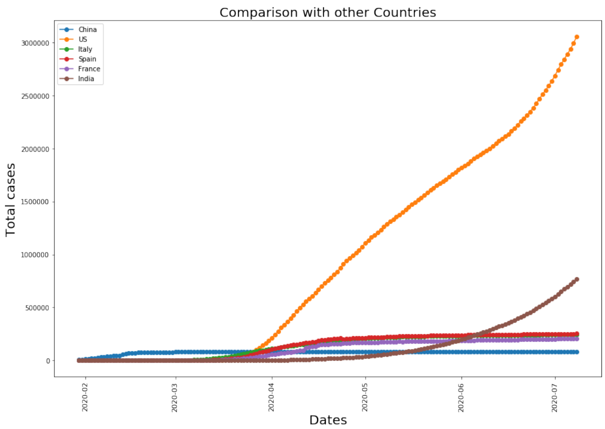

Figure 2: Covid-19 Spread Comparison with other countries

2.2 Gender Analysis

We can see from the 3 that Covid-19 is having more spread on men than in females.It may be because in India it is more common that the work involved with more exposure is done by men.

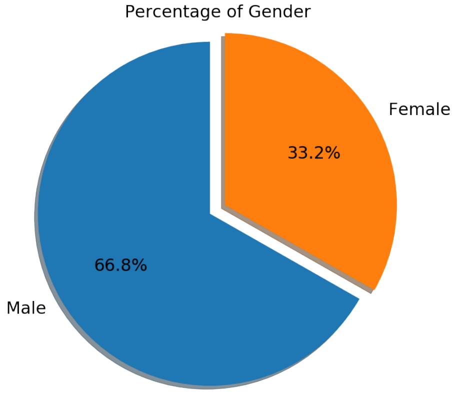

Figure 3: Covid-19 GenderWise Distribution of Confirmed Cases

2.3 State wise Analysis

From 4 we can see that Maharashtra is the most affected state in India followed by Tamilnadu, Delhi and Gujarat.

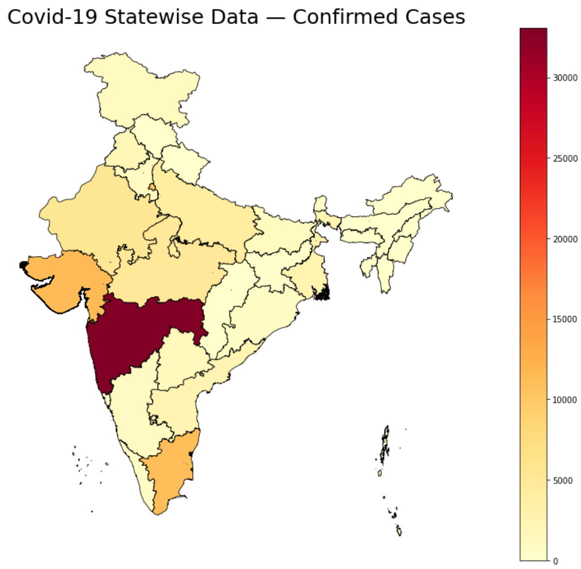

Figure 4: Covid-19 Statewise Data of Confirmed Cases

2.4 Network Analysis

Having knowledge about network science has led us to model and analyse this data in the from a large network. Using various graph properties on this network for interpolating and extracting new information. We viewed the data as collection of nodes and edges with the nodes being the patients and the edges being the spread of virus caused by them. Our edge list maps the patient ID to their contacts. Here the contacts could be the patients spouses or someone close. They also could be travelling from different countries to India. Here in the image 5 below we can see a network with different origins of COVID-19 and the large red circle resembles spread caused by that particular patient. We considered only the patients at the initial stages of the virus in our experiment. This consists approximately 200 nodes.[] From this graph network we can

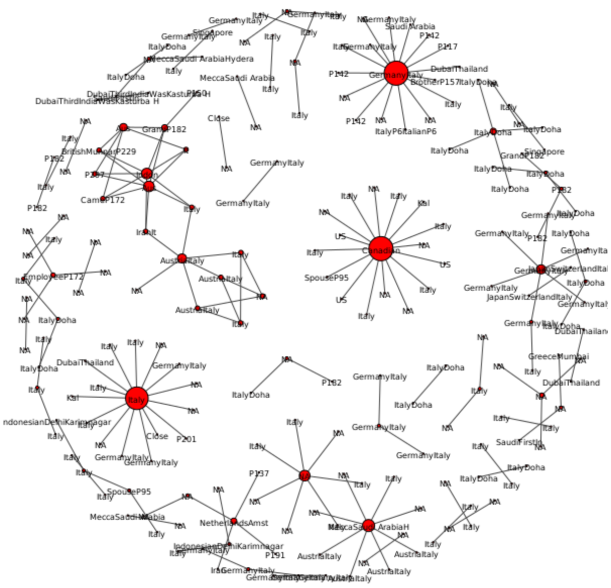

Figure 5: COVID-19 Network with limited nodes

find the betweenness of the nodes the thicker lines in the below image 6 represent higher betweenness. This centrality measure quantifies the number of times a node (i.e patient) acts as a bridge along the shortest path between two other patients. This helps us find the most used path by the virus to enhance the spread. Now we can find it and isolate it. Similarly 7 the betweenness of the nodes help to identify the patients who need to be isolated to reduce the spread of the virus. The bigger nodes are the ones that need to be isolated to effectively contain the spread of virus. We were also able to detect many communities being formed at the initial stages of the virus. In 8 all the similar color nodes belong to same community or cluster. This helps us to find the patient zero in each individual cluster separately and understand the virus more easily. We also see that in further stages as the

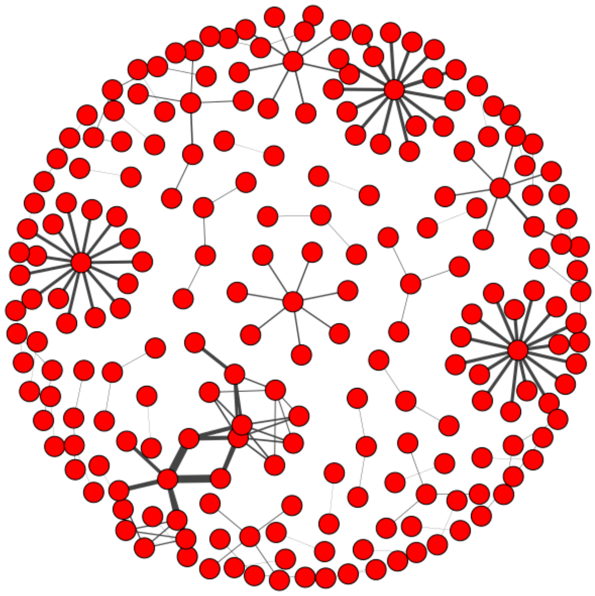

Figure 6: Betweenness of nodes

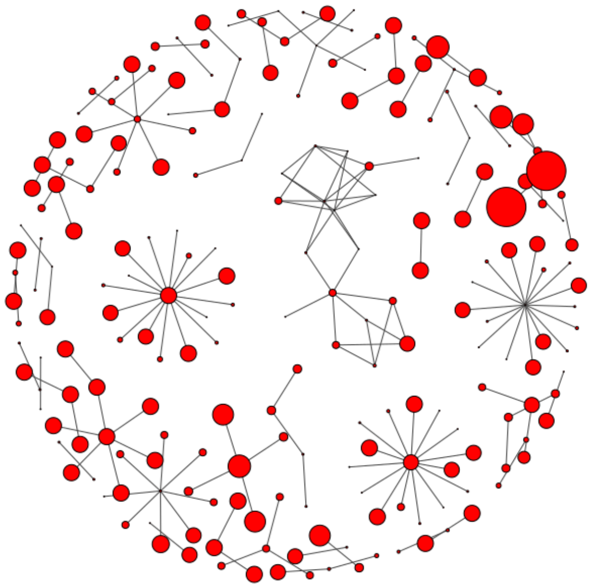

Figure 7: Betweenness of edges

virus evolves it is because of the cross community spread. By finding the k-core from the network we can single out the largest k-core encountered. This measure can help us identify small interlinked core areas on a network. We see a 2-core in the below image 9. This is the core area that can increase the rate of further spread of the virus. This 2-core needs to be isolated from other cluster groups to contain the spread to minimal. By plotting the largest cliques we can find out the patients that require our immediate attention so as to

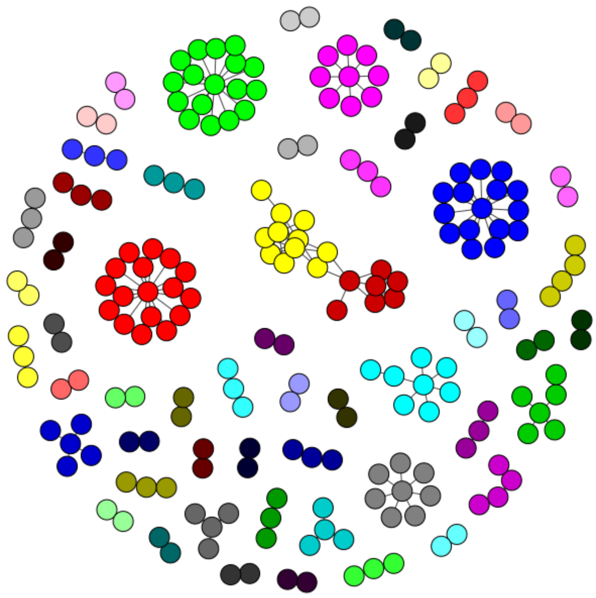

Figure 8: Various communities detected in the network

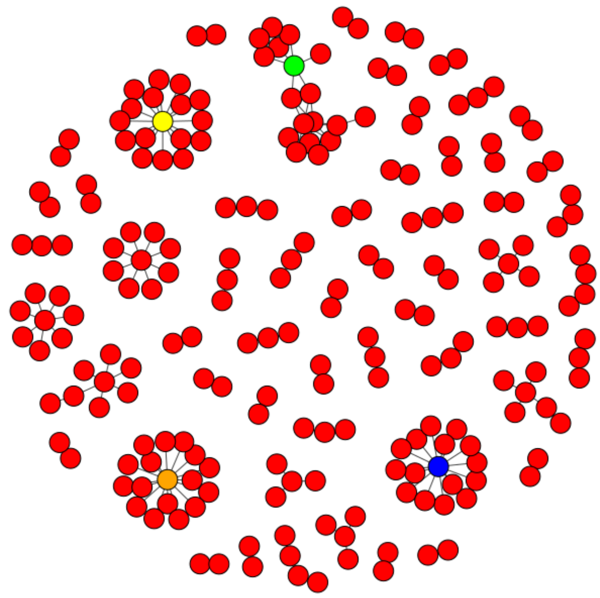

Figure 10: Largest cliques detected

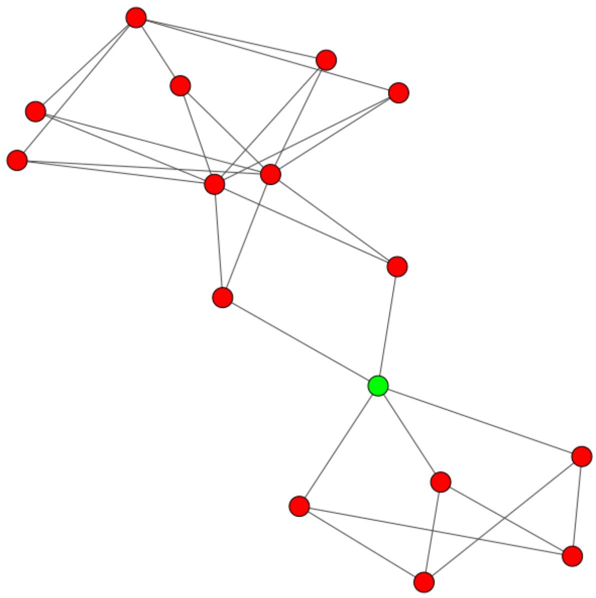

Figure 9: Largest k-core encountered

isolate them and treat them quickly. From 10 we see different coloured nodes that are responsible for the bigger clusters around them and these are the ones that needed to be taken to the emergency health care centers and kept on ventilators as soon as possible.

3 Modeling

3.1 SIR Modeling

An SIR model is an epidemiological model that computes the theoretical number of people infected with a contagious illness in a closed population over time.

3.1.1 Vanilla SIR

Our aim here is to model the spread of the disease. One of the simplest ways to do so is using a compartmental model. Basically, these models divide the population into different compartments and with other parameters like probability of spread between people in different compartments, model the spread of the disease in the population. One of the simplest compartmental models is SIR model. It stands for Susceptible - Infected - Recovered model. This model divides the population into three compartments:

- Susceptible (can still be infected, but healthy at the moment)
- Infected
- Recovered

This is in no way a complete picture since it doesn't include the probability of dearth from the disease and other important factors like the incubation period. We will look into this in more detail later. For the purpose of this study, we will try to understand the effectiveness of this model by trying to model the spread of Covid 19 in India. SIR model uses the following parameters :

- Total population - N
- Initial number of infected and recovered - I_0, R_0
- Beta parameter β - The number of people an infected person infects per day.
- Gamma γ - Reciprocal of number of days an infected person spreads the disease.

Note that the parameter beta β can be defined as product of r_0 and gamma.

$$\beta = r_0 * \gamma$$

Where r_0 is the basic reproduction number of an infection. To model the spread in India, we take the statistics available as of 1st April from official sources and try to model the spread in the next 90 days i.e for roughly three months. We use a r_0 value of 1.3, with initial infected and recovered counts being 1837, 160. Now for gamma γ, we assume once a person has been diagnosed he will be quarantined and hence will not be able to spread the disease anymore. From the available data, on average a person with Covid will spread it to others for 5 days before being diagnosed. So γ will be $1/5$. Prediction for the next 90 days is shown in 11

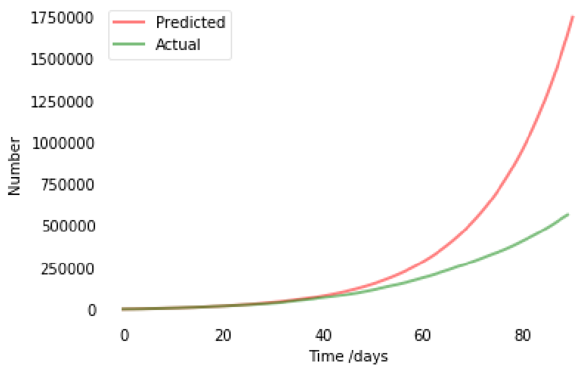

Figure 11: The prediction of SIR model for Covid spread in India for 90 days starting April 1st.

It can be seen that the predictions are a bit off. This is because this model is not entirely representative of different compartments that can exist. This has more effect as the duration increases. This cannot be used to model the spread for longer duration. But the predictions are pretty accurate for shorter periods, and given it's simplicity, it can be used to model the spread for short periods. These insights can be useful for decision making.

3.1.2 Extensions to SIR - SIER model

We include more compartments in the SIR model. Some extensions could be :

- Dead state - people can die after contracting the disease.
- Exposed state - People don't spread the disease immediately after contracting it. There is an incubation period before which they can start spreading the disease and a second incubation period before they start spreading the disease and start showing symptoms which is when they get diagnosed. Exposed state is the state in which a person has been exposed but hasn't started spreading yet.

The dearth rate of Covid is low and in India, it is as low as 3%. So we don't expect this compartment to be a major factor but we still consider it for completion. The other compartment i.e the Exposed compartment is a huge game changer. This is important because, in the standard SIR model we assume a person starts spreading the disease immediately after being infected which is not the case in reality. There is a gap before a person can do that and this gap means the spread doesn't happen as fast as predicted by SIR model. This is the reason why SIR model predicts such high numbers as duration increases. To accommodate these two compartments, we bring in three more parameters into the model:

- Alpha α - dearth rate (0.03 in India)
- Rho ρ - rate of dearth: Reciprocal of number of days before an infected person dies
- Delta δ - exposed incubation period - The period during which a person is exposed but can't spread the disease yet.

This model is similar to [5] We try to model the spread in India from April 1 for 90 days again, this time with the new compartments. From the data we take ρ as $1/15$ and δ as $1/3$ i.e three days before an infected person can start spreading. Note that these are average values and will be sufficient here since we are modeling the spread as a whole and not at an individual level. Results shown in 12

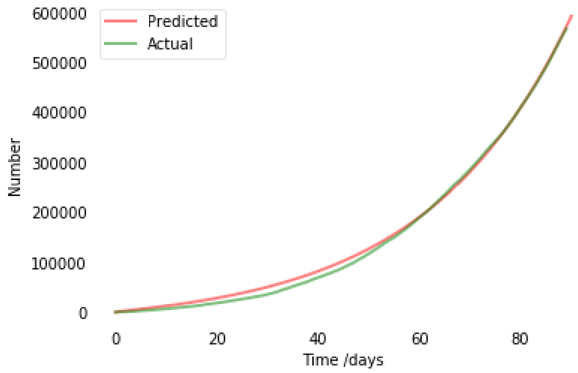

Figure 12: The prediction of SIER model for Covid spread in India for 90 days starting April 1st with additional compartments, Dead and Exposed

The prediction is pretty accurate now. We can see how the Exposed compartment has made a huge difference. Here, we have used the average r_0 value of 1.3. In reality, in India r_0 kept decreasing over time and it cannot be treated as a constant.

3.1.3 Time dependent r_0 values

Until now, we have assumed r_0 is constant throughout the spread but this is not true. r_0 varies with time, spread and actions taken like quarantine, lock downs etc.. For example implementing the social distancing rules, movement restrictions and curfews will bring down the r_0. So if we could predict the r_0 value in the coming days, the spread could be modeled accurately. This can also be used for decision making purposes. For example, the government can analyze how a measure that is going to be taken will effect the spread in the coming days (ease lock down? open stores? etc). We can see this using the r_0 values from the available data and seeing what the SIER model predicts.

We use a r_0 value gradually decreasing as shown in 13. Previously we have been using the average value of 1.3.

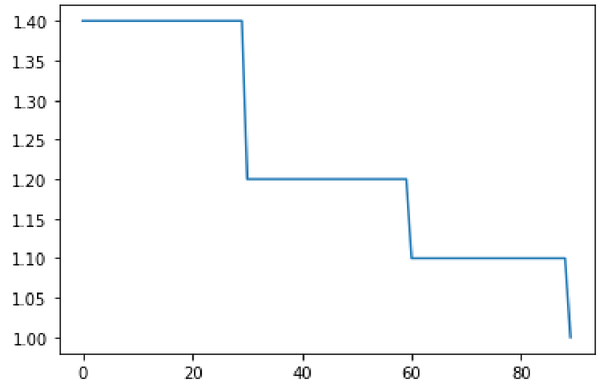

Figure 13: Time dependent r_0 value gradually decreasing over 90 days.

The results are shown in 14.

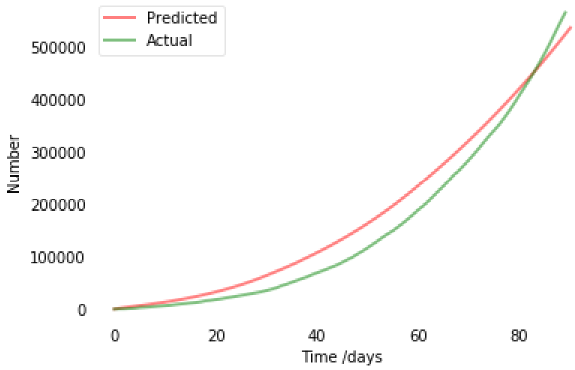

Figure 14: The prediction of SIER model for Covid spread in India for 90 days starting April 1st with time dependent r_0 values.

3.1.4 More improvements to SIER

The predictions in 14 might seem a bit inaccurate than the previous prediction, but decreasing r_0 will make the predictions more stable over time and are representative of conditions in real world. We can include more factors to make the predictions more stable :

- Resource and age-dependent fatality rates that will enable us to model overcrowded hospitals, populations with lots of young people etc.
- Demographic details - state divisions and populations, cities and rural area information etc can be used.

These will also allow us to model the spread at an individual level.

3.2 Prophet Modeling

Prophet is an open source library by Facebook[6]. It is based on decomposable models. The three main model components are trend, seasonality and holidays. It provides us with the ability to make time series analysis. Prophet uses simple intuitive parameters and has support for including impact of custom seasonality and holidays which in our case is lock down. Prophet Uses time as a regressor, and tries to fit several linear and non linear functions of time as components. It is in fact modeling seasonality as an additive component and in effect, framing the main forecasting problem as a curve-fitting problem rather than looking explicitly at the time based dependence of each observation within a time series. A trend may consists of Saturating growth which can be defined as a varying capacity C(t) for the time series forecasts or Change points where an unknown or unforeseen may fall upon in that specific time frame and it may as well affect the forecast. As we allow the number of change points to increase it introduces flexibility in the model but we are also faced with two problems overfitting and underfitting. The model also forecasts the effects of seasonality modeled by fourier series. Fourier order defines whether or not high frequency changes are allowed to be modelled. Holidays and events incur predictable shocks to a time series. So these are considered separately and additional parameters are fitted to model the effect of holidays and events. Now We use this model to do a time series analysis on the data for predicting the total number of COVID-19 cases for the next 30 days. We take a number of parameters such as period of lock down, current number of cases, deaths etc into account for fitting the model. The major task here is forecasting of total cases in INDIA for next 30 days. The model sees the trend in data weekly and during the holidays to help in predicting the future cases. From this we can also see how the lock down is affecting the current situation. By observing the trend in holidays which in our case is lock down period we can fairly state how much longer the lock down needs to be extended in order for the cases to minimize drastically. [7] The model forecasts the cases in India for the next 30 days. One such day values are highlighted in the 15. As we can see that the cases are increasing day by day. We took into account the lock down parameter as start date being on 2020-03-24 and the end date as 2020-05-31.(According to lock down in India on 2020-05-17). As we can see in the

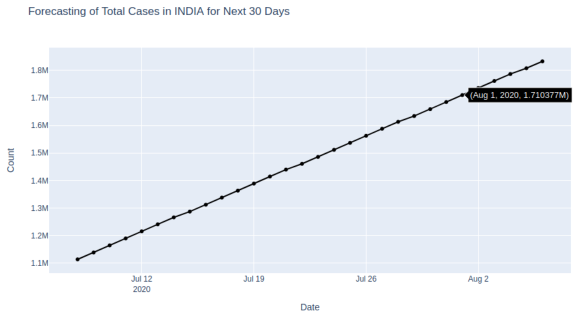

Figure 15: Predicted Cases

image 16 the actual cases are almost coinciding with the predicted number of cases. The

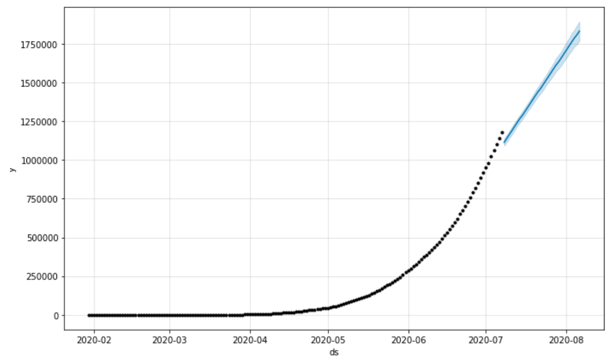

Figure 16: Actual Cases

model also forecasts the deaths in India for the next 30 days. One such day values are highlighted in the 17. As we can see that the death toll also increasing day by day. As we can see in the image 18 the actual deaths are coinciding with the predicted number of deaths.

3.2.1 Training Parameters

We look into how the training parameters vary by observing the trends from 19 weekly and on holidays. We look into how the training parameters vary by observing the trends from 20 weekly and on holidays.

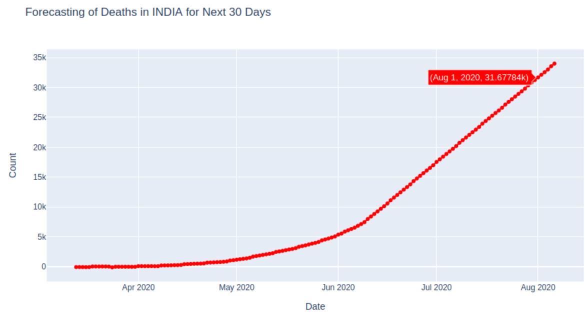

Figure 17: Predicted Deaths

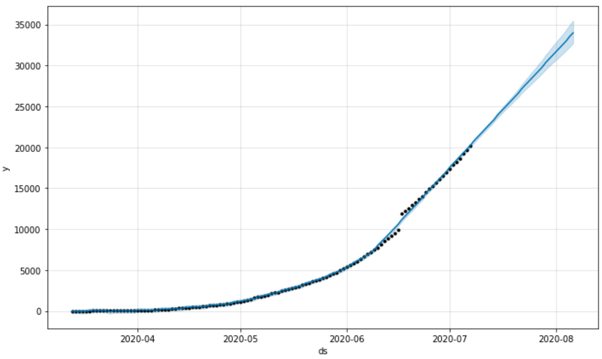

Figure 18: Actual Deaths

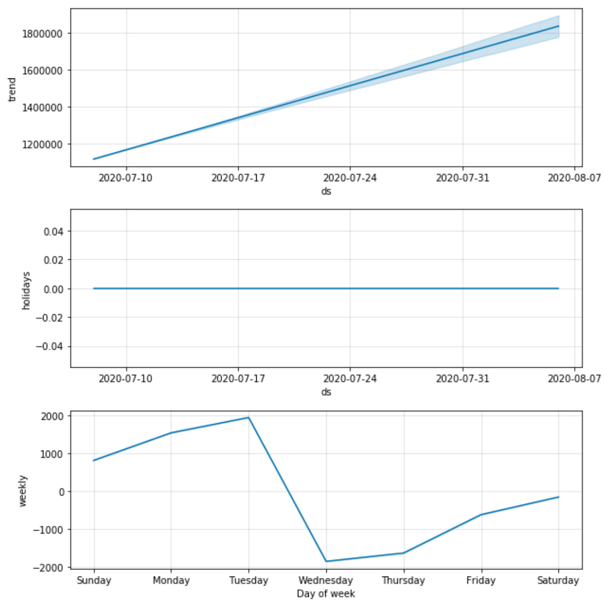

Figure 19: Training parameters for cases

3.2.2 Performance Metrics

We calculate the Mean absolute percentage error (MAPE). MAPE is an metric used to determine the success of a regression analysis. We see how this metric varies as the number of days keep increasing finally on the 30th day we have mape value of 0.759806 from 21.

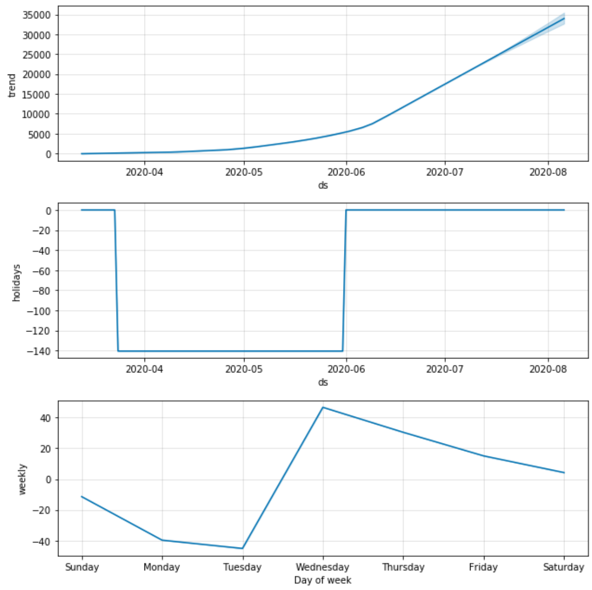

Figure 20: Training parameters for deaths

Finally it keeps deceasing. We calculate the

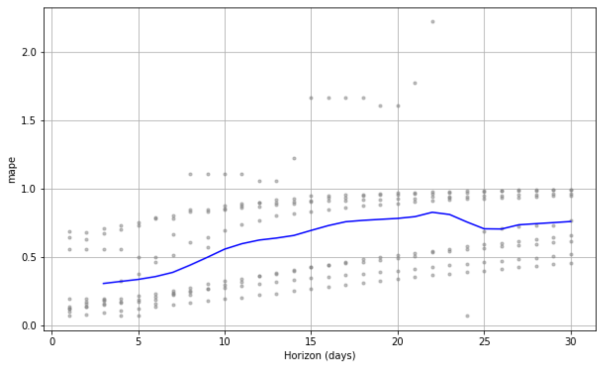

Figure 21: Performance metrics for cases

Mean absolute percentage error (MAPE) for the evaluating the model for deaths as well. We see how this metric varies as the number of days keep increasing finally on the 30th day we have mape value of 0.587360 from 22. Finally it becomes constant.

4 Conclusions

This work explores the current situation of Covid-19 in India and provides basic insights on the data(as of on 30th June 2020). Various models are build using the different parameters such as initial transmission rate, population of India, duration of the lock down etc..The prophet model is able to predict the number of future cases with an mean absolute

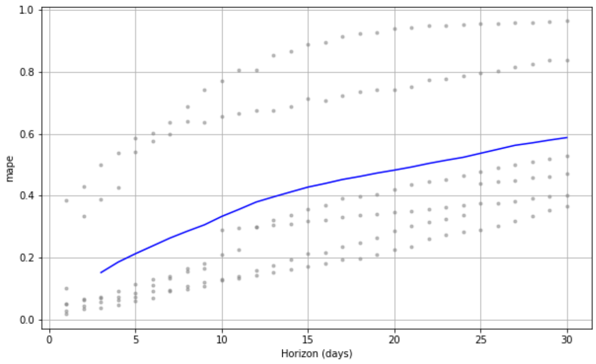

Figure 22: Performance metrics for deaths

percentage of 87.78%. The model works for only short term periods i.e about for 1 to 2 months. But the model is not well suited for long term analytics since it doesn't take into consideration the population cap of a which is an essential characteristic of epidemiological model. SIR model along with its extensions addresses this issue but additional parameters have to be involved and to predict accurately in long term, a diverse set of parameters are to be consider such as isolation, and how will different age groups interact with and without lock down and the diffusion between different states and cities etc.

5 Future Work

As part of future work age structures contacts can be included into modelling so that it improves the overall results in estimating the spread of the virus and duration of lock down period. We can also look at how other models work for extended prediction of number of cases. Network analysis with more number of nodes for performing further interpolation on data is also required due to the high velocity and volume of data being generated every day. To look at other parameters such as movement between the states and how good is the quarantine etc to estimate better results should also be include into the model.

Combining the epidemiological properties of SIR model with Prophet model which is very accurate, we can see if we will be able to make better predictions.

References

[1] A.G. McKendrick W.O. Kermack. Contribution to mathematical theory of epidemics, p. roy., soc. lond. a mat.,115,, 1927.

[2] Umme R. Siddiqi Arpita Howlader Samrat K. Dey, Md. Mahbubur Rahman. Analyzing the epidemiological models, outbreak of covid-19: A visual exploratory data analysis approach, 2020.

[3] Data from india health ministry website:. https://www.mohfw.gov.in/.

[4] Ravi Suggala and Anurag. A survey on prediction and detection of epidemic diseases outbreaks. 03 2019.

[5] Chih-Cheng Sheng Bin, Gengxin Sun. Spread of infectious disease modeling and analysis of different factors on spread of infectious disease based on cellular automata, 2019.

[6] Letham B. Taylor SJ. Forecasting at scale. peerj preprints 5:e3190v2, 2017.

[7] Ankit Choudhary and analyticsvidhya. analyticsvidhya. Generate Quick and Accurate Time Series Forecasts using Facebook's Prophet (with Python R codes).