COVID-19 in California: Predicting Cases and Deaths

Authors

- Pranav Bhasin
- Muskaan Goyal

Introduction

The World Health Organization (WHO) issued a “public health emergency of international concern,” on January 30th. In less than 5 months, COVID-19 has spread across the world infecting over 4.53 million people, and causing nearly 307K deaths in over 212 countries and territories. This disease, from the class of influenza-like viruses like MERS and SARS, is a keen reminder of the Spanish Flu pandemic.

The COVID-19 pandemic continues to have a devastating effect on the health and well-being of the global population. A critical step in the fight against COVID-19 is predicting the number of new cases and deaths. As most countries are following some form of social distancing measures, an important consideration for governments in lockdown is the growing number of cases and deaths despite these measures. This motivated our team to model the spread of the disease in the form of new cases and deaths. By leveraging time-series datasets and population, we would like to predict the number of new cases and deaths in California.

Research Question(s)

- How can we predict the number of new cases of COVID-19 in California using the time series data of new cases in all states in the United States and their respective populations?
- How can we predict the number of deaths due COVID-19 in California using the time series data of deaths in all states in the United States and their respective populations?

Description of Data

For this project, we used the following datasets:

- **time_series_covid19_deaths_US.csv**: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv

- **time_series_covid19_confirmed_US.csv**: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv

Exploratory Data Analysis, Data Visualisations and Data Cleaning

For answering our research question, we needed to focus on the state-level data rather than counties. As the data present in the tables **time_series_covid19_confirmed_US.csv** and **time_series_covid19_deaths_US.csv** is presented on a county-level, we needed to process our tables to get the required data. However, before we could do that, we needed to address the null values, and values that don't fit our criterion of being a state.

Missing Values

While trying to address the null values, we noticed that in both the tables there are 7 Admin2 values and 4 FIPS values which are missing. From the documentation of the datasets we knew that Admin2 and FIPS represent unique counties. We had also checked if all the rows belonged to the US country and found a few wrong entries but luckily these coincided with the 5 missing entries of Admin2. We noticed the following observations after researching and exploring:

- We found that counties **Dukes and Nantucket** and **Kansas City** have missing FIPS values for historic reasons. We combined these counties with their respective states data because we felt these counties exhibit the similar mainland characteristics like their respective states. We replaced these missing values with the value “Unassigned” and included the numbers belonging to these two rows in our dataset and model.
- We also noticed that the 5 missing entries of Admin2 are **US Overseas territories**. As they are islands, they might not accurately represent characteristics similar to that of the mainland states. So we decided to remove these territories from our dataset and did not consider them in our predicting model.
- We noticed that the other two missing entries of Admin2 column belonged to the province states Province States titled **Diamond Princess** and **Grand Princess.** They are in fact cruise ships that were stranded in the seas for a long-time. As the people on board were disembarked from the ship, these people are already being counted in their respective states. So, we can safely remove them from our count.
- The last two of the missing entries in FIPS are of **correctional facilities**. On further researching these numbers, we couldn't find if these were already included in the count for Michigan. Nonetheless, as these facilities are locked and are not

accessible openly by the people outside, we are excluding these numbers from our dataset and model.

Preparing The Dataset

After replacing and removing the missing values in both our datasets, we dropped the columns - "UID", "iso2", "iso3", "code3", "FIPS", "Admin2", "Country_Region","Lat", "Long_" and "Combined_Key" as they did not provide us with useful information on the state for the model. We also made sure that both the datasets consist of the same set of the states in the United States. We utilised the population column in the death data set to create another data frame which represented the current population of different states in the United States.

Data Visualisations

While epidemics often grow in an exponential fashion, from the heatmap we can see how the values in the different states were extremely low initially, and grew significantly toward the end supporting our notion. To account for the same in our linear model, we will linearize out data by normalization.

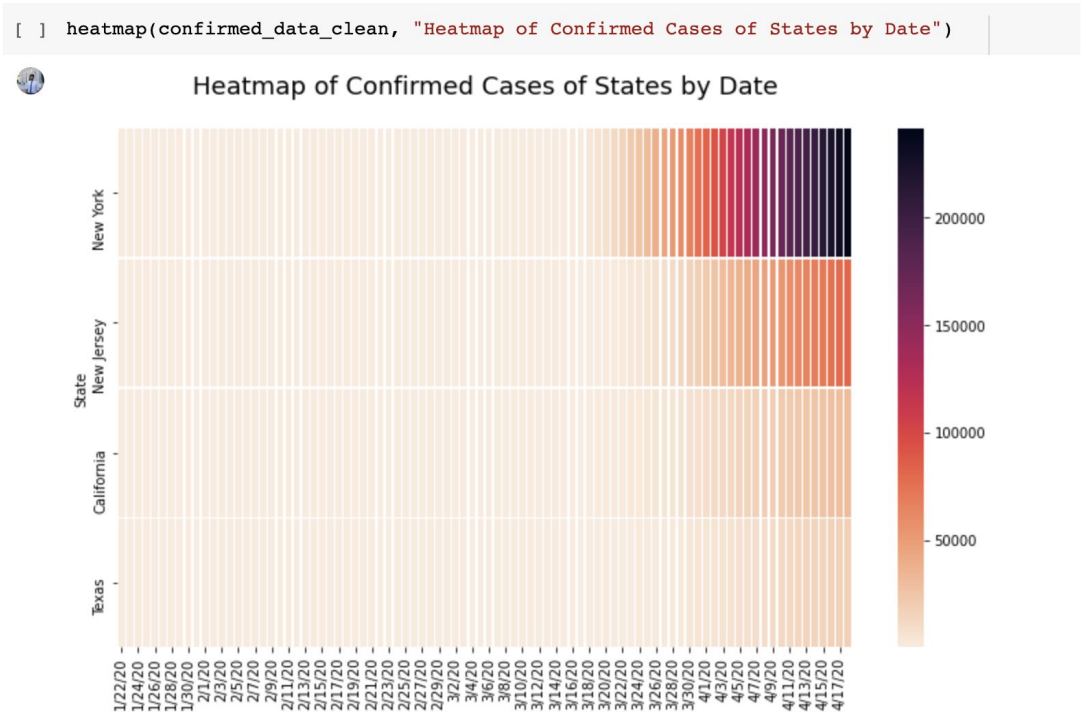


Figure 1: Heatmap of Confirmed Cases by Date

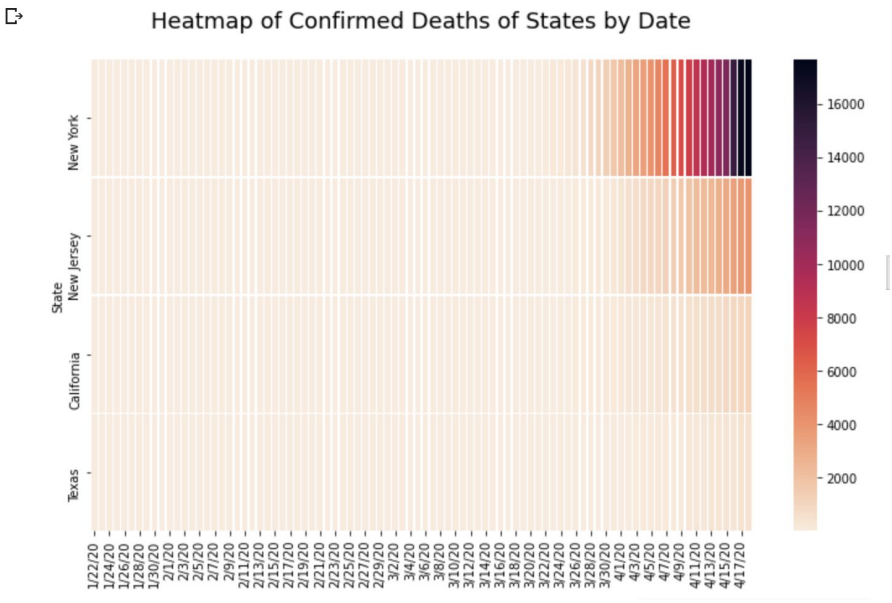


Figure 2: Heatmap of Confirmed Deaths of States by Date

Visualization Plot Of Fraction Of People Infected

We can utilise these visualisations to further analyse total cases of covid 19 in the different states in the US in terms of relative growth curves and fraction of total population.

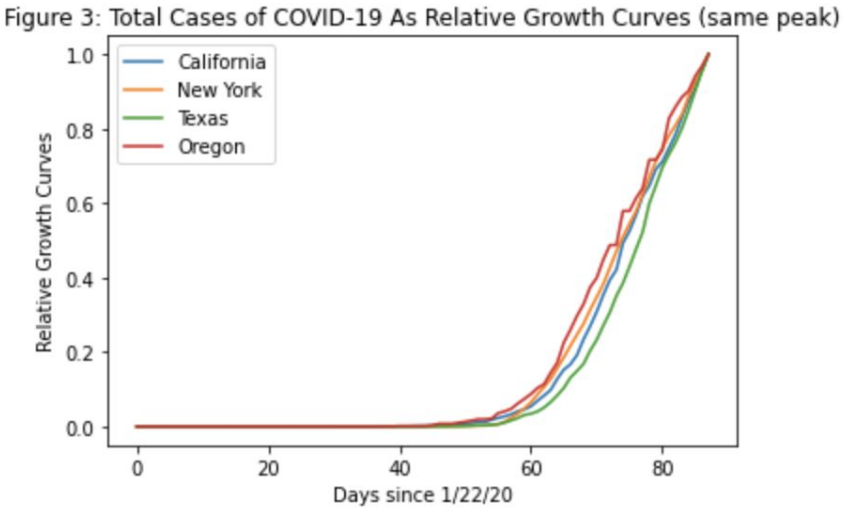


Figure 3: Total Cases of Coronavirus Cases as Relative Growth Curves

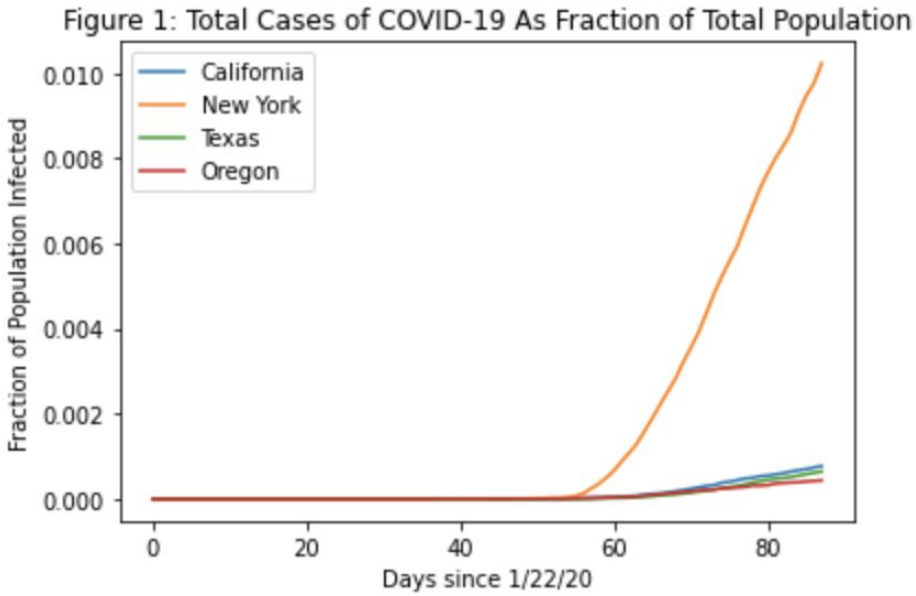


Figure 4: Total Cases of Coronavirus Cases as Fraction of Total Population

Description of Methods

Normalisation:

After cleaning and preparing our data, we decided to normalize the data as different states have different populations. We will first divide all the entries in both the datasets by their respective state populations to help us represent the number of cases and deaths on the same scale. Essentially, we will be predicting the number of coronavirus cases and deaths in terms of the fraction of total population of the region/state. The goal of normalisation of the data was to change the values of the numeric columns in the dataset to a common scale without distorting differences in the ranges of values in the dataset.

Reverse Normalisation:

We decided to create a reverse normalisation method so that we could transform the predicted number of coronavirus cases and deaths which is represented in terms of the fraction of the region's population to the standard number of cases and deaths. The methods normalisation and reverse normalisation helped us better calculate the loss of our model without distorting the differences in the ranges of the values existing in our datasets.

Logarithmic Transformation and Exponential Transformation:

We also created methods for logarithmic and exponential transformation for future use to help us project the exponentially growing number of coronavirus cases into a linear space which will help us improve the accuracy of our regression classifier.

Generate Training and Validation Data Set:

We utilised the above methods to create a method “get_Data” which helps us generate clean and preprocessed training and validation data sets. We also utilised the following two parameters to create our training set and validation data set:

- **Threshold**: Only take data from countries after the country has reached the threshold number of cases or deaths. This allows us to filter out the many data points that would be entirely zero.
- **Window Size**: How many preceding days we are using for datapoint. E.g. If we want to use the last 4 days as a prediction for the 5th day, window size would be 4.

Model and Assumptions:

After generating our training and validation sets for both the datasets, we created a prediction model of the number of coronavirus cases in a region, specifically California in this case which depends on the historical trend of coronavirus cases in the region as well as other states in the United States. To build this model, we went through the following iterations:

First Iteration:

- **Ridge Regression**: We utilised the ridge regression module of the sklearn library to build a regression classifier for our prediction model. We trained our ridge regression model using our training data. We decided to use ridge regression over lasso regression because we were working with fewer features and did not help with feature selection. Ridge regression also helped us avoid overfitting the model and helped us estimate reasonable reliable approximations to true population values.
- Assumptions: To utilise ridge regression we assumed linearity, constant variance, no outliers and independence.
- Result: After training our dataset and playing around with different alpha values, we found that the training accuracy is 0.9374 and validation accuracy is 0.667 for the alpha value equal to 0.002.

After playing around with alpha values, we realised that we can optimise the alpha value to improve our training accuracy and validation accuracy.

Second Iteration:

- **RidgeCV:** We also utilised the ridgeCV model of the sklearn library to improve our regression classifier by adding multiple alpha values as its parameter. We decided to try RidgeCV model because we thought it would help us find an optimum alpha value among a few different alpha values which results in the best training accuracy and validation accuracy among those values.
- Assumptions: This model also has the similar assumptions as the previous model - linearity, no outliers, independence and constant variance.
- Results: After training our dataset and playing around with different alpha values, we found that our training accuracy improved from 0.937 to 0.975 and validation accuracy improved from 0.667 to 0.868 when we reduced our alpha value from 0.002 to 0.001.

  After exploring these two models, we realised that we can significantly improve our prediction model by optimising the parameters of our model - threshold, window size and alpha. This gave us an idea to utilise ridge regression and optimise the parameters using tuning to improve our model.

Final Iteration:

- **Ridge Regression + Optimisation using Tuning:** After the two iterations, we decided to optimise our model using tuning where we first optimised our threshold value while keeping the other two parameters constant. For the number of coronavirus cases, we found the optimum threshold value did not change and remained as the value one. We then utilised this optimised threshold value and alpha as zero, to find our optimized window size to be 29. We further utilised these optimised values to find the optimum value of alpha to be 1e-15. We utilised the same tuning method to optimise the parameters for the model predicting the values of number of deaths in california. We believe this model has the same assumptions as the previous iterations.

Summary of Results

Initially we utilised the model Ridge Regression model in sklearn library which helped us train our data. We got decent training accuracy and validation accuracy, but we realised we could do better. After playing with different alpha values we decided to iterate on our model using Ridge Regression with Cross Validation model in sklearn library. Through the Ridge Regression with Cross Validation model we were able to improve our model by improving the alpha value. This iteration helped us realise that we can improve our model by improving the parameters of our model - threshold, window size and alpha value. In our final iteration, we utilised regression model and optimisation using tuning to improve our model's training accuracy. After tuning the parameters for the number of coronavirus cases in california, we noticed that our training loss decreased from 5009 to 892, validation loss decreased from 4066 to 190, training accuracy increased from 0.96 to 0.999 and validating accuracy increased from 0.75 to 0.999. After tuning the parameters for the number of deaths in california, we noticed that our training loss decreased from 252 to 19.9, validation loss decreased from 148 to 4.2, training accuracy increased from 0.729 to 0.996 and validating accuracy increased from 0.58 to 0.998. Therefore we can say that optimising these parameters significantly improved our accuracy rates and decreased our losses in both training and validation datasets.

Prediction Model of Number of Coronavirus Cases In California:

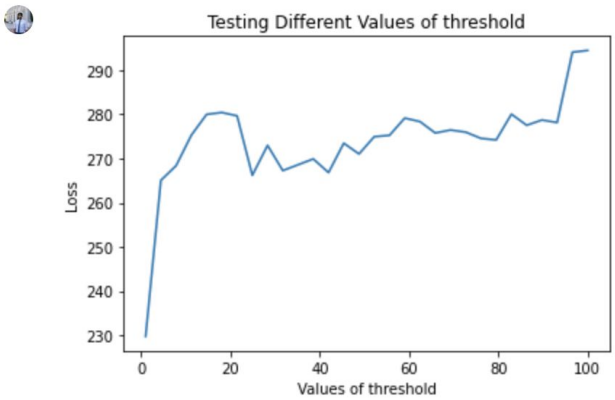


Figure 5: Line Plot of Threshold Values vs. Loss Function Value

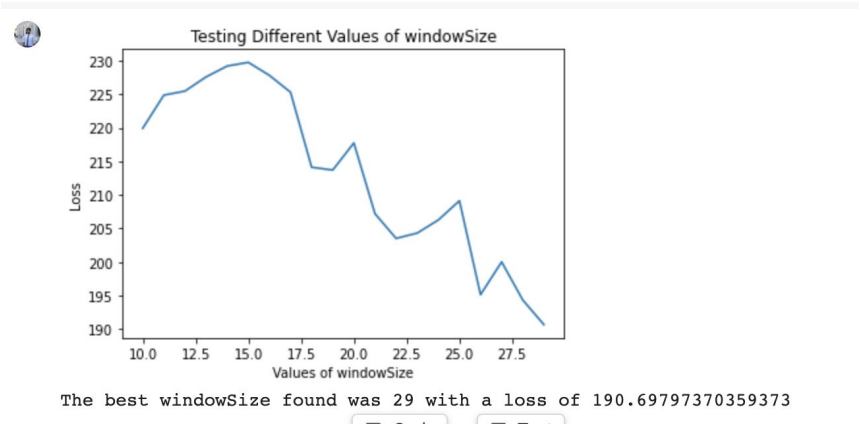


Figure 6: Line Plot of Window Size Values vs. Loss Function Value

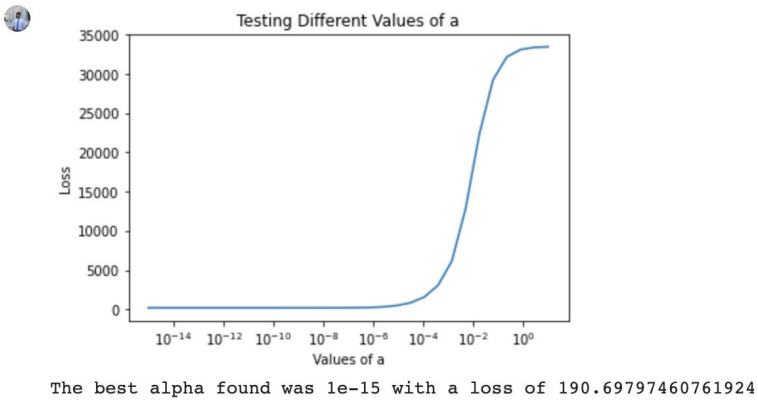


Figure 7: Line Plot of Alpha Values vs. Loss Function Value

Prediction Model of Number of Coronavirus Deaths In California:

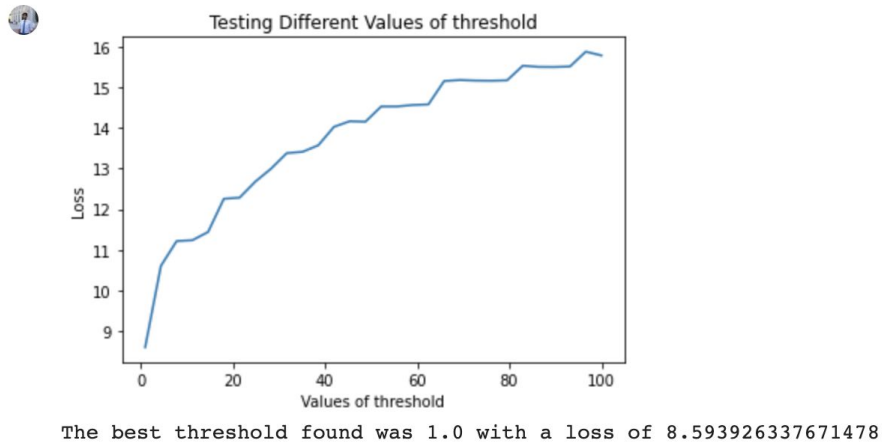
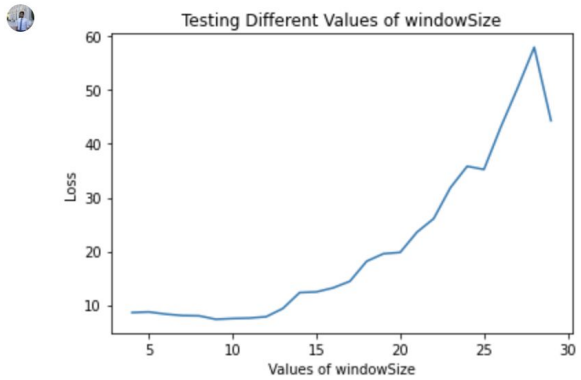
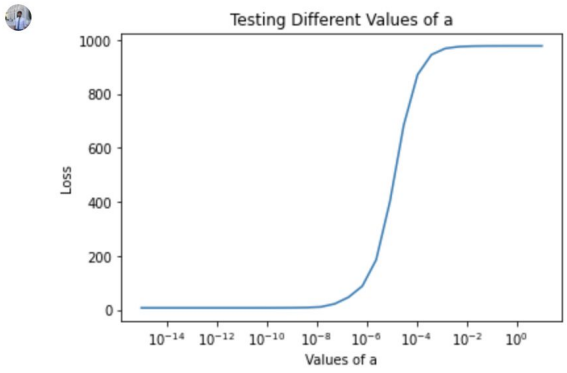


Figure 9: Line Plot of Threshold Values vs. Loss Function Value




```
The best windowSize found was 9 with a loss of 7.335347878160615
```

Figure 10: Line Plot of Window Size Values vs. Loss Function Value




```
The best alpha found was 1e-15 with a loss of 7.335348588158796
```

Figure 11: Line Plot of Alpha Values vs. Loss Function Value

Discussion

- What were two or three of the most interesting features you came across for your particular question?
    - The most interesting features we came across our particular research question were the historical trends, window size parameter and population. It was also interesting to note the exponential growth in the time series data. It was specifically interesting to see how the model's accuracy increased with different window sizes. We did not expect that the optimal window size would be 29 for predicting the number of coronavirus cases. The safety guidelines and features also looked interesting and useful but we could not utilise them.

- Describe one feature you thought would be useful, but turned out to be ineffective.
    - When we were optimising our parameters, we expected the optimised threshold parameter to be different than one. We expected the optimised threshold parameter to significantly decrease the loss but in reality, the optimised value of the threshold parameter turned out to be ineffective.
- What challenges did you find with your data? Where did you get stuck?
    - Initially, when using Ridge regression, we encountered negative values of model score. While we were reducing loss function value, and achieving nearly 90% accuracy on the test set, we weren't able to fix this issue. On further research, we learned more about the limitations of the model, and improved it by hyperparameter tuning of the alpha value.
- What are some limitations of the analysis that you did?
    - Our project was focused on getting the best results possible with the time series data. This model currently takes only into consideration the State Population and Time Series data. This means that the model is susceptible to changes. As social distancing measure changes, this model would deviate from the initial conditions. However, by limiting the window of training data to last 'n' days, we can still achieve a reasonable model accuracy.
- What assumptions did you make that could prove to be incorrect?
    - One of our assumptions was that most of the countries are following social distancing in some form or the other. However, as countries slowly ease into economic activities, this assumption could prove to be incorrect. However, as discussed above, it is relatively easy to adapt to this situation.
- What ethical dilemmas did you face with this data?
    - Nothing in particular. The data is reasonably anonymized to protect any and all individuals. One
- What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?
    - Reliable data about state-wide recovery rates would've helped us further explore the problem through SIR modelling. Without this data, we couldn't approach this avenue. We tried numerous approaches to approximate this data, however, we weren't able to convince ourselves of those approaches.
- What ethical concerns might you encounter in studying this problem? How might you address those concerns?
    - To solve the problem of prediction of new cases, one might consider contact tracing technologies to gather more data. With the use of such

data, contact data collected using low energy bluetooth handshakes provides more concrete evidence of potential exposure to the virus. However, this approach poses numerous privacy concerns as the technology exposes the locations of individuals that opt-in to be a part of the experiment.

Future Scope

- This model currently takes only into consideration the State Population and Time Series data. This means that the model is susceptible to changes. As social distancing measure changes, this model would deviate from the initial conditions.
- This model can be improved by considering additional features, including changes in social distancing norms.

Resources

Data Cleaning and Data Preparing

- **http://www.ds100.org/sp20/syllabus/**

Time Series and Forecasting

- **https://www.analyticsvidhya.com/blog/2019/12/6-powerful-feature-engineering-techniques-time-series/**

We would also like to thank our amazing professors, GSIs, course staff and fellow classmates for a great semester.