Covid-19 Virus Spreading Forecasts in the U.S

Tina Teng, Truong Le
May 13, 2020

Abstract

The novel coronavirus COVID-19 identified in December 2019 was declared by WHO on March 11th as a global pandemic. On April 18th, the epidemic of SARS-CoV-2 caused a total of 2,330,764 cases and resulted in 162,894 deaths globally. The COVID-19 virus has attracted numerous researchers' attention to explore treatments as well as statistical information to prevent its spread. This research project utilizes several well-known public data as an attempt to analyze the spreading statistics of the COVID-19 Virus, specifically in the U.S. The project applies a handful of practical data science tools and approaches taught in Data100 at UC Berkeley to predict the spread-reduction of the virus in the near future.

Introduction

The outbreak of coronavirus disease has created a global health crisis and deeply impacted people’s lives. Its rate of contagion and patterns of transmission have threatened social security and financial livelihoods of many workers and families. It also motivates renowned researchers and companies to invest a huge amount of money into finding vaccines and treatments to quickly put an end to the disease. One productive approach to accelerate this process is to use statistical analysis, embracing modelling, parameter estimation, and hypothesis testing to bridge the gap between mathematical theory and public health practice. Those statistical analyses provide guidance to control the spread of the pandemic, given that they could produce accurate predictions about the virus into the near future. This project aims to explore three fundamental approaches to forecast the future trends of the COVID-19 virus, including: Log Transformation, Linear Regression Modelling, and Epidemiological SIR Modelling.

Description of Data

We will concentrate specifically on the statistics of the COVID-19 in the U.S because it is the country that has been severely affected by the virus. We focus on the quantitative data, including death, confirmed, and recovery cases as they are the indicators of the future spread of the pandemic. We also pay much attention to the daily changes in those data because they can

identify the “curve” of the pandemic. Lastly, we also attempt to observe the differences between different states in the U.S to get a sense of how each state handles the pandemic and when each state can open to operate as normal.

Description of Methods

Question
We focus primarily on how to predict the spread of the COVID-19 virus in the near future, how long will this situation go on, and when it will come to a close.

Exploratory Data analysis (EDA)
We used several EDA functions, including read_csv, head(), shape, info(), describe, value_counts() to conduct preliminary EDA and get a glimpse of how our data look like. After understanding the data definitions, shape, info, and description of the data, we checked how many NaN values are in the data set for data cleaning.

Data cleaning
As we observed from our EDA, there are a lot of NaN Values in our data frames. Thus, we applied some data cleaning strategies in order to make our data frames in an easy-to-model format:

1. There are many NaN values in our data frames, after careful considerations, we decided to fill these missing values with 0s. There are two reasons we think that this is the best choice to do:

   Firstly, we observed that the `confirmed_data` and `deaths_data` have only 11 missing values (in FIPS and Admin2). These values do not relate to any models or predictions we are going to use. Thus, we will drop these two columns.

   Secondly, for unavailable data in `counties_data` and `states_data`, we will fill them with 0s and just use available data to compare our predictions with data from confirmed.csv and death.csv. In other words, we will not use these features in `states_data` to train our models because the NaN values may negatively impact our analysis in some mysterious ways.

2. In `counties_data`, we see that some rows have all 0.0s values, so we drop these rows. In this data frame, we also see that only 2 rows have all non-zeros values as shown above. Thus, we should not filter out any more value since it will not help our modellings.
3. We can see from part 1.E above that `states_data` has more than 1 country, including: 'US', 'Canada', 'United Kingdom', 'China', 'Netherlands', 'Australia', 'Denmark', 'France'. Thus, we want to take into account only the U.S by filtering out all other countries. We also dropped some rows that do not directly relate to the states of the U.S, for example: Recovered row,

Samoa 3th row, Diamond Princess 9th row, Grand Princess 13th row, Puerto Rico (Caribbean island and unincorporated U.S. territory).

Visualizations

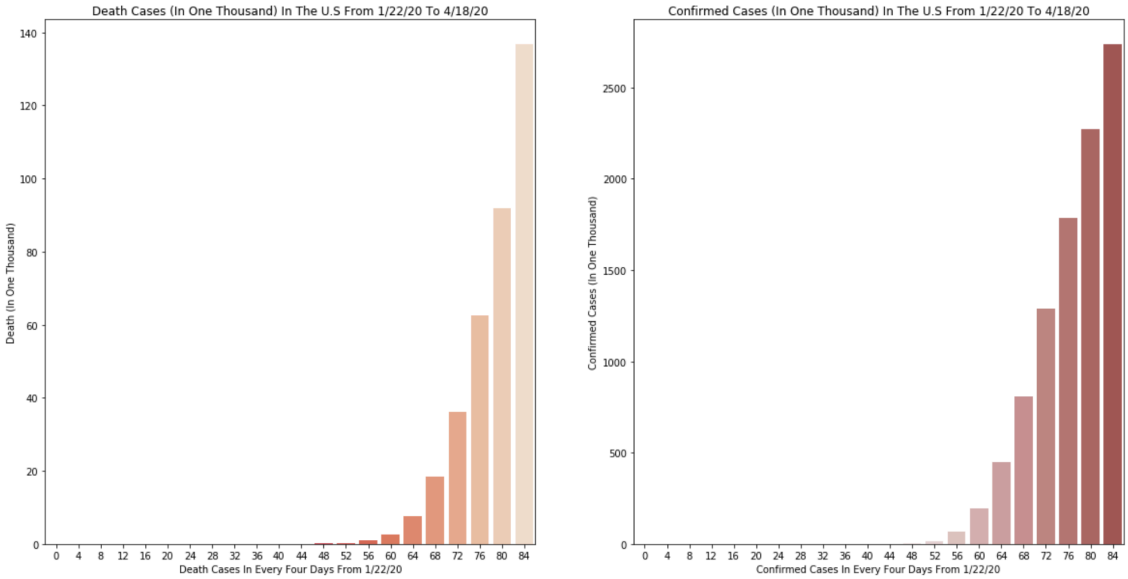


Figure 1: Comparing the Total of Confirmed vs Death Cases in The U.S

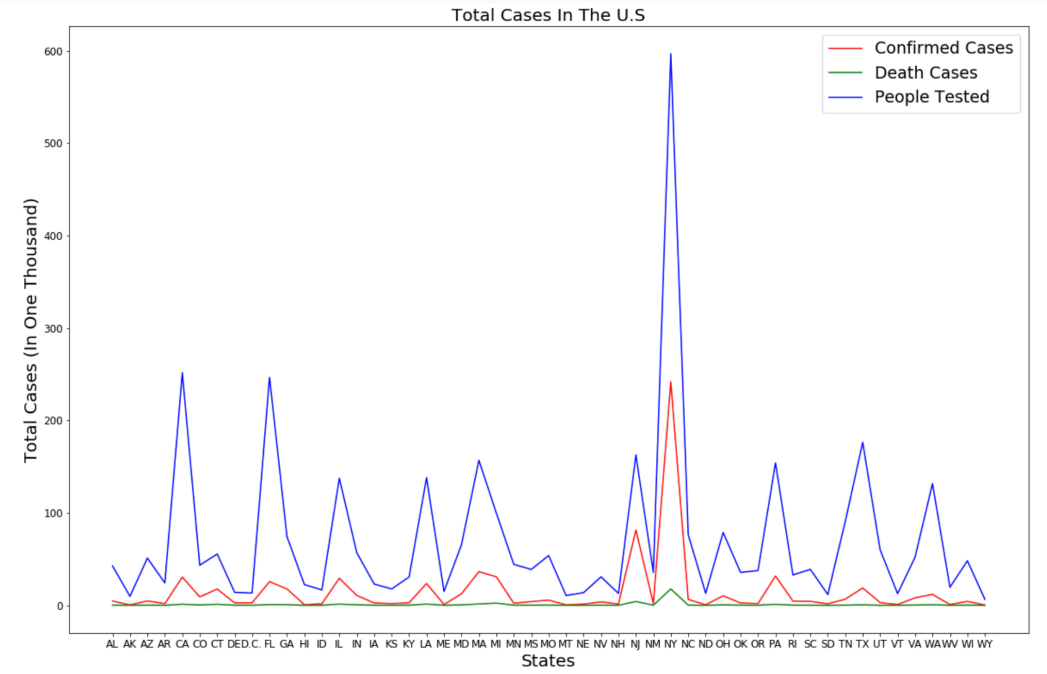


Figure 2: Comparing the Total of Confirmed Cases vs Tested-Cases in the U.S

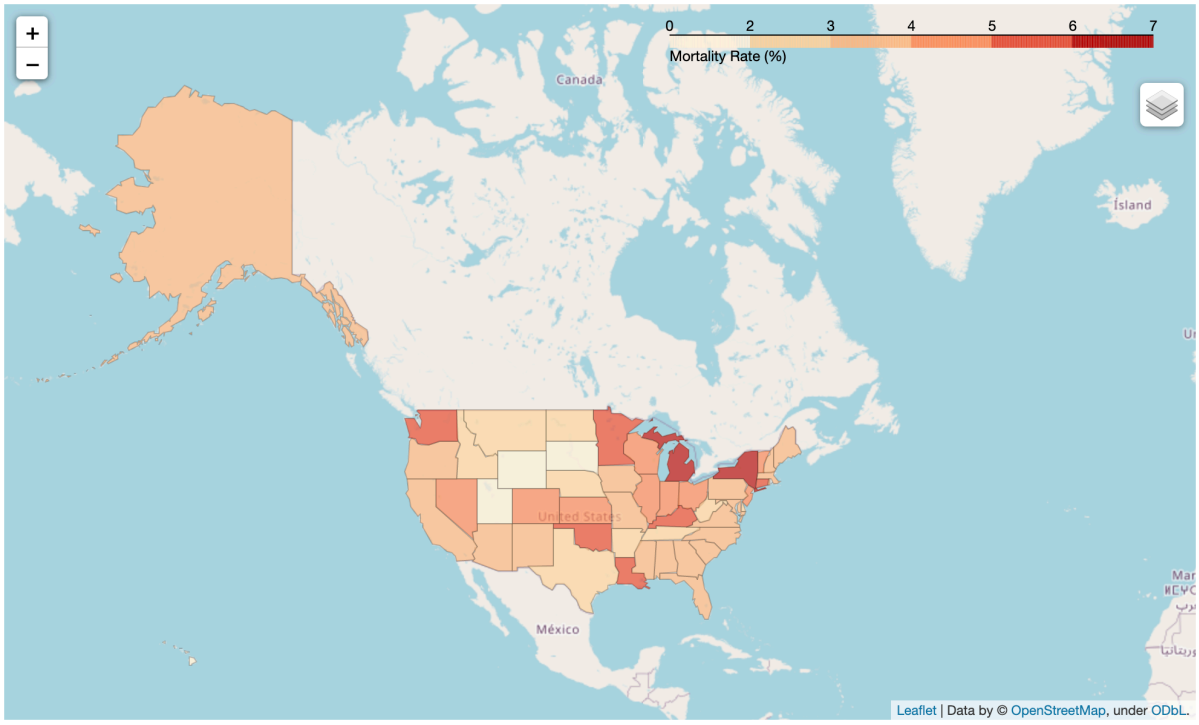


Figure 3: Geographical Mortality Rates in the U.S

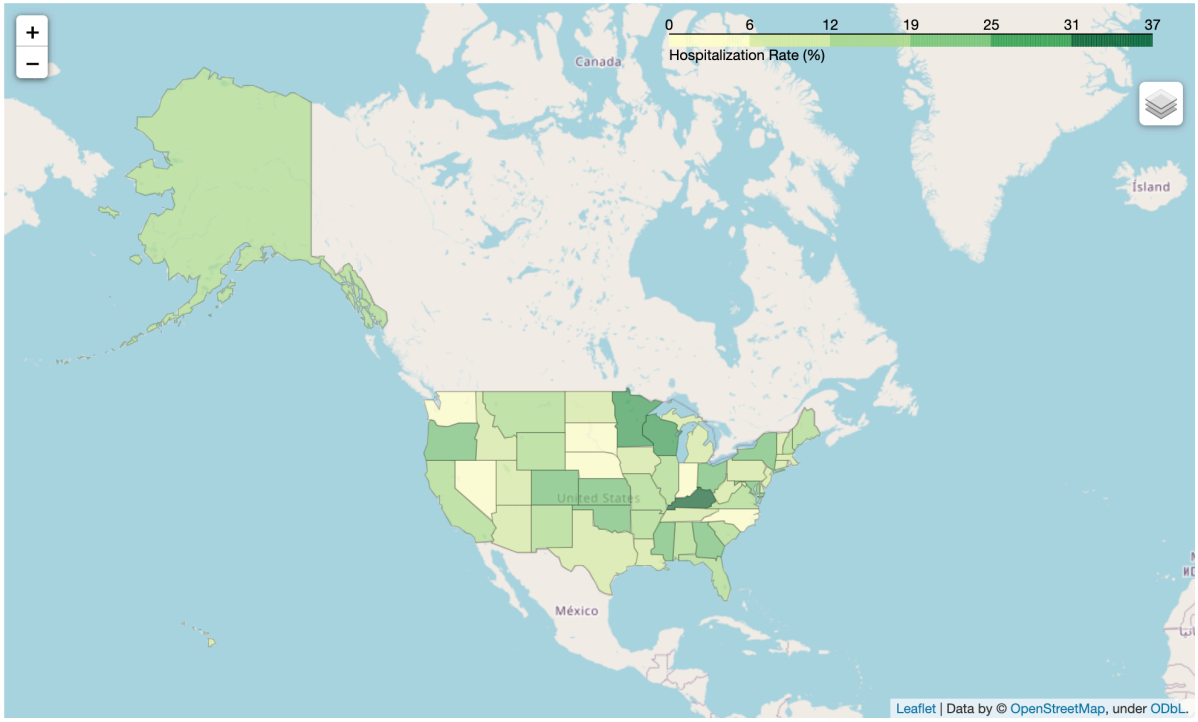


Figure 4: Geographical Hospitalization Rates in the U.S

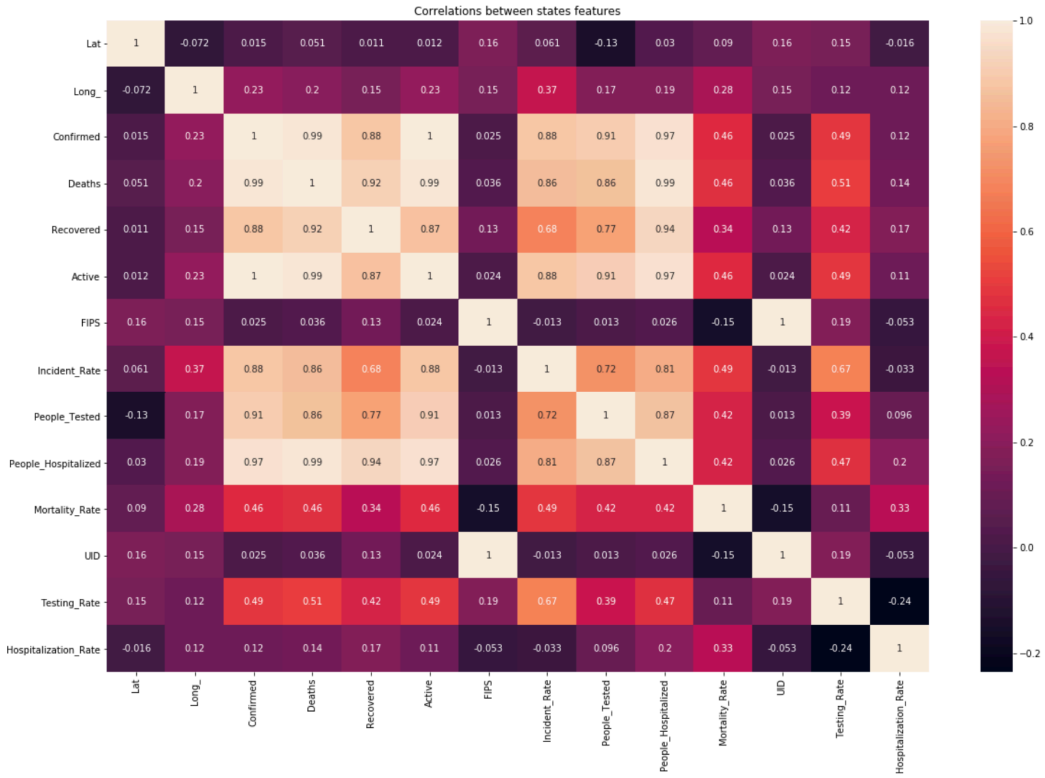


Figure 5: Correlations between features in all states in the U.S

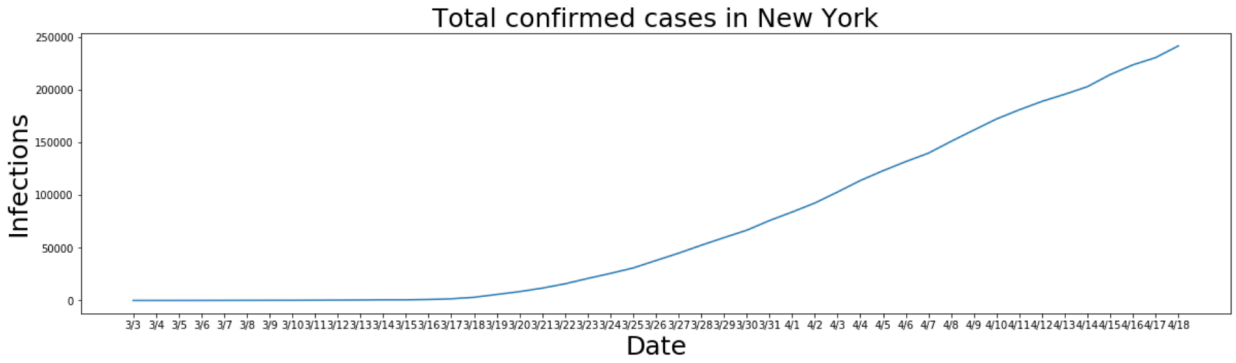


Figure 6: Coronavirus Growth in New York from confirmed_data dataframe

Methods

Experiment#1: (The Naive Experiment)

The first model we created was linear regression using exponential growth to predict future pandemic growth. In the process of trying to visualize all the possible data to predict the future of the pandemic, we figured a simple Linear Regression Model that may be able to predict the cases in the future. This model uses logarithms transformation and the given statsmodels.api package to predict the pandemic growth in the future. Since a linear regression is not suitable for extrapolating exponential trends, using log transformation will allow our algorithm to better interpret and process the data. We have our formula derived in the Jupyter notebook.

Experiment#2: (The Accurate Experiment)
This model calculates the combination of the daily increase of 2 variables: Confirmed cases and Death cases using the diff() function in order to linearly predict the future trends of the pandemic. This diff() function is used to calculate the difference of a series element compared with another element in the previous row. To make it simple, we assume there is no bias towards our calculation. For example, we assume everyday in the week has no anomalies occur, but in fact, weekends usually have a smaller or larger number of new test cases, depending on the regions and number of test kits available. These combination of death and confirmed cases will be our feature to predict new cases in the daily future.

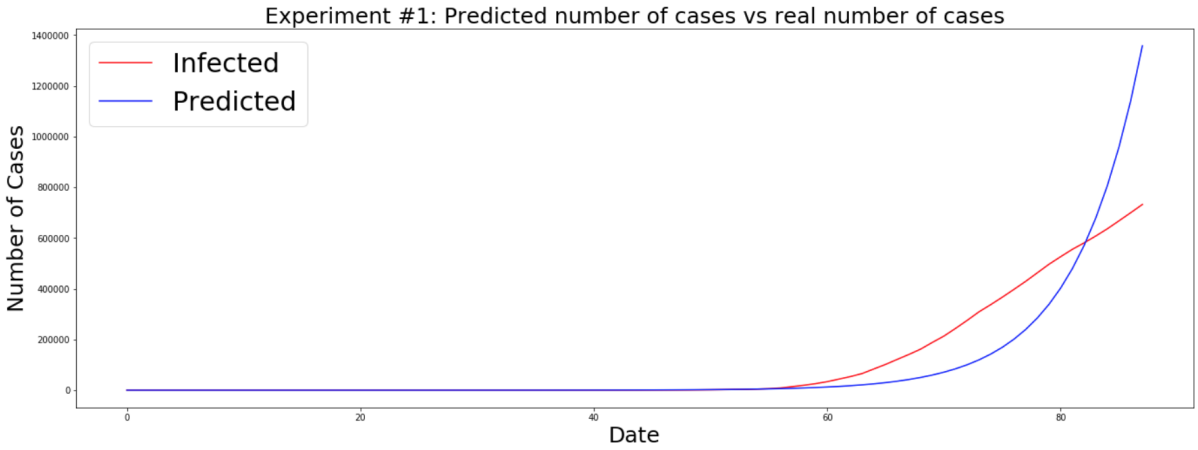
Experiment#3: (The Modern Experiment)
The last model attempts to use the well-known model for estimating the dynamics of the COVID-19 virus, called SIR Model. The model intends to divide the virus growth rate of the population into 3 segments: Susceptible, Infected, and Recovered. By using the fundamental tool of calculus – derivatives, and the 4 key parameters that are derived and explained in the Jupyter notebook, we can forecast the three segments. We adapt our implementations largely from the detailed note on SIR modelling on Piazza.

Statement of the Model and Assumptions used for Inference
Experiment#1: (The Naive Experiment)
Statement of the model: Using Log Transformation to transform an exponential growth function to a linear function to predict the spread of COVID-19.
Assumption of the experiment: we can simply model the exponential growth of the pandemic by using a single constant growth factor and keeping track of the recorded data; chances of reinfection (virus rebounds) are close to zero for all people.

Experiment#2: (The Accurate Experiment)
Statement of the model: Using the Daily increases in Confirmed and Death Cases to accurately predict the future trends of COVID-19.
Assumption of this experiment: the combined values of the total confirmed cases and total death cases can be a reliable variable for estimating the current epidemic progression, and for modelling future trends.

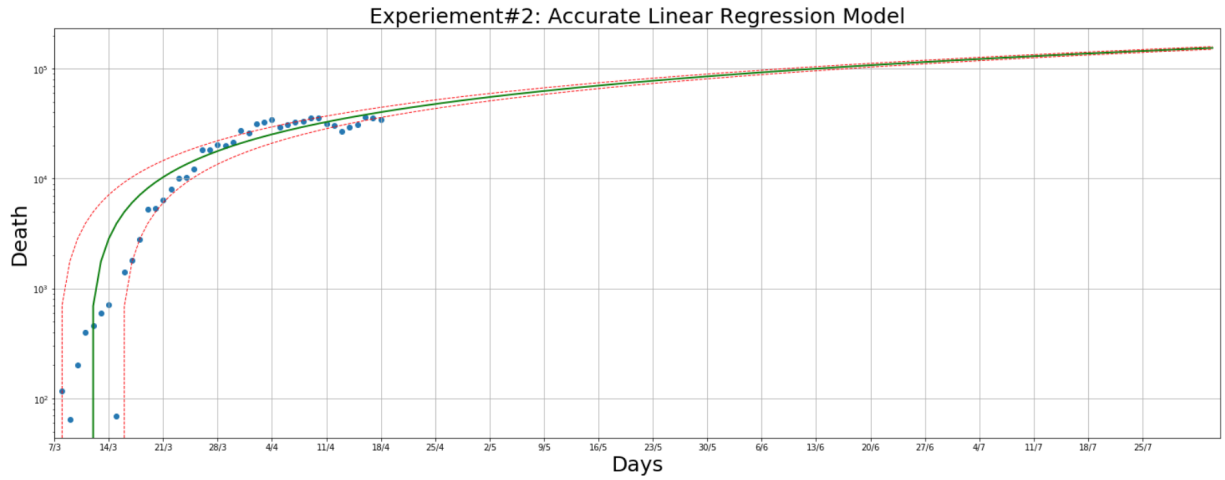
Experiment#3: (The Modern Experiment)
Statement of the model: Separate the population into 3 groups in order to calculate the growth rate of COVID-19 (Compartmental model)
Assumption of this experiment: the population is fixed at a certain time of our experiment; all other factors such as age, sex, race, social and financial status do not affect the probability of being infected; each person in the population can only be in 1 group described above.

Summary of Results



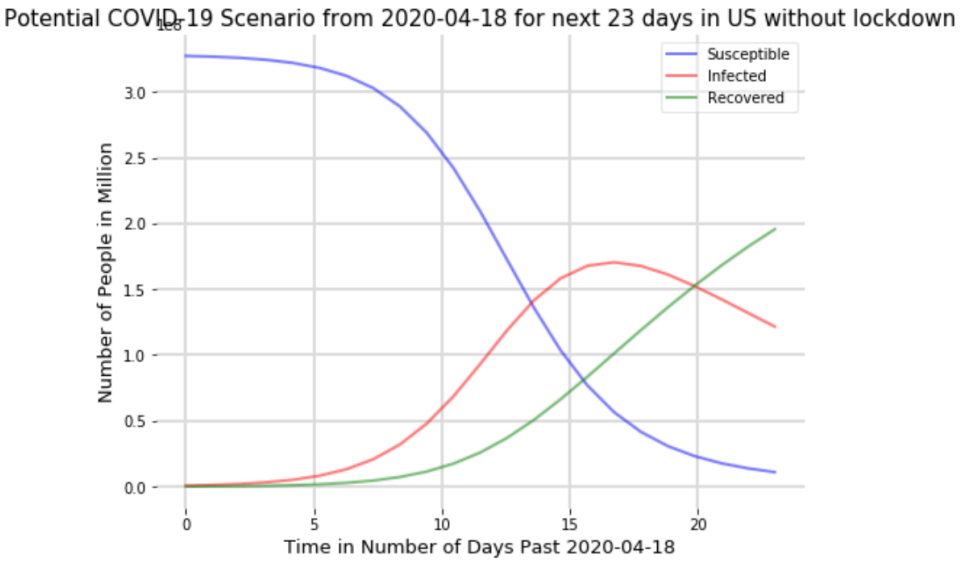
As we can see, the Exponential Growth will only fit the epidemic at some points in the timeframe given. The reasons for this might be:

- At some point, the infected people will be at its peak and thus the Linear Model does not accurately predict the result. For example, on 3/10/20, the number of people infected is only half the number of people predected. On contrary, at the last day in the above data frame (3/21/20), we have 25725 infected, but can only predict half the number based on our formula.
- At some other point, healed people will not spread the virus anymore and when (almost) everyone is or has been infected, the growth will stop. At another point, infected people who did not know they're infected will affect their surrounding people.

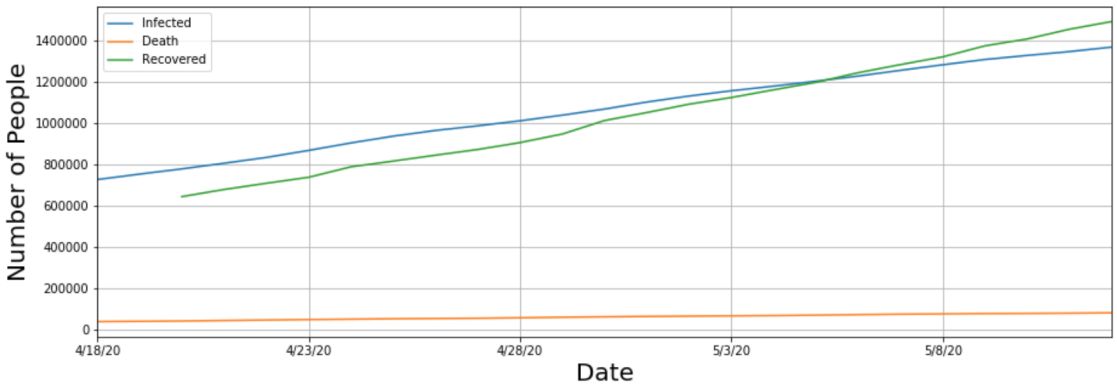


We compared this Death Prediction of the U.S with the actualy death of the U.S now (~80K) according to WorldOmeter website, this turns out to be very accurate as we can see from the graph above since 10^5 = 100K people.

As we make our Linear Regression Model, we see that the coefficient of determination R^2 of the prediction (the proportion of the variance in the dependent variable that is predictable from the independent variable) is ~90%, which is very high for this complicated epidemic. Moreoever, our RMSE is only ~4000 cases, which is very low compared to our first experiment.



The graph above displays the potential COVID-19 scenario from 2020-04-18 for the following 23 days till 2020-05-12 in US without lockdown. Given that most of US in lockdown during that time period, we wanted to observe the data from 2020-04-18 till 2020-05-12.



The SIR model data showed that the number of infected would increase to a little over 160 million around May 4th and decrease afterwards. However, the data plotted based on the historical US trends indicate that the infected increased at a less steeper slope and continued to increase beyond May 4th. Despite so, it fortunately remained well below 160 million within the given time frame. The SIR model data further shows that the number of recovered increased up to 200 million by May 12th. The number of recovered in historical US trends increased to a little over 1.4 million. The discrepancy may be due to the fact that less people were infected than the model predicted. Thus, we can say that shelter-in-place and lockdown of US has helped prevent the infected case from reaching the predicted potential thus far.

Discussion

FEATURES

Most Interesting Features

We found the daily increases in confirmed, recovered, and death cases can represent precisely the trends of the virus in the future.

We also found that the dynamics of COVID-19 can also be inspected using the SIR model with the same 3 simple features/characteristics above.

Ineffective Feature

In our first experiment, we tried to use logarithm transformation learned in CS70 to transform an exponential function to a linear form so that we could create a Linear Regression Modelling, but it turned out that the actual COVID-19 graph was much more complicated than a simple exponential growth function. Although the "log transformation" method has been known to be the most popular approach to transform skewed data to normal distribution, it does not seem to be useful in this complicated case because there are a lot of factors we need to take into consideration to model this pandemic case.

CHALLENGES

Since the given data have many NaN values, it was sometimes difficult for us to interpret them in our data analysis. For example, the recovery column in states_data has many NaN values. And we believe if this feature had more complete data, it could be easier for us to utilize this feature for our modelling.

LIMITATIONS

Although our quantitative method of training models makes the data very consistent, precise, and relatively easy to analyze, we saw that it may not be robust enough to explain the complex issues of the COVID-19. For example, people with underlying health issues, children and elderlies with compromised immunity, and even race can also be a big factor to model our data. Another important factor not detailed in the data are the political parties and policies that may affect each states' decisions in dealing with the virus.

For the reason above, our analysis was based largely on the assumptions of confirmed, deaths, and recovery cases. It may not capture all necessary factors to train more complex models. In other words, we believe that our Linear Regression Models may oversimplify this intricate COVID-19 pandemic, and we need to take into account other numerous external factors.

ETHICAL DILEMMAS & CONCERNS

Ethical Dilemmas

An ethical dilemma faced was with the selection of data and how the extrapolation could make an impact on the findings. Imagine if we had selected data from Taiwan instead of the US. As

Taiwan has not implemented the same strictness of social distancing and lockdown procedures, yet the number of infections and deaths are greatly under control, the extrapolated findings could lead to recommendations and policies that are applied to the rest of the world. However, the US and other countries would greatly benefit from stricter social distancing and self-isolation. As our question was focused specifically for the US, it was only appropriate to use the data from the US despite the limited data it has as the outbreak started later than the rest of the world.

Ethical Concerns

Ethical concerns surround data analytics, from algorithmic fairness to societal consequences. As explained previously, if our findings will affect public policy and the public opinion, it could negatively impact the lives of others if the data was not applicable. While our models are based on the current data in the US, our models are biased by our selection of data from certain time periods and are limited in terms of features and factors we considered. Unfortunately, there may be no concrete ways to address these concerns given the constant evolving nature of the pandemic. Fortunately, there are more and more research being conducted about COVID-19 and public health officials and policy makers could rely on multiple studies in order to make balanced, well-informed decisions.

LIMITATIONS

Strengthening analysis

As we stated above, the additional data could be political dominations of each state, the economic loss, etc. These features could help us to make more Logistic Regression Model to address our question on how to better predict the future spread of COVID-19.

Evaluation of approach and limitations of the methods used

Our first naive approach, using log transformation, has a lot of limitations. The reason is because it can only be used to model a simple exponential growth, as it only has a simple growth factor b. Whereas Coronavirus is an extremely complex and intricate problem. For example, there are more than just a single independent variable, such as population density, probability of interaction of each region, the political domination of certain states, etc.

Our approach to train our model in experiment #2 can accurately predict the spread of COVID-19 in the future. It has RMSE of only 4,000 (compared to 112,000 in our first experiment) and its accuracy score is 90%. As a result, we can see that our model was improved significantly.

Our third approach uses a well-known and widely used model epidemic model. We believe the limitation of the model is that it ignores random effects that could contribute significantly to the growth rate of the virus. For example, the early stages of each region is largely different, thus creating discrepancies between different regions.

SURPRISING DISCOVERIES

It took only 13 days for New York to have a total of people infected increase from 0 to 100,000. Even after nearly 2 months of the 'Shelter-In-Place' Order, the daily new confirmed cases keep developing, it reaches ~25,000 new cases in the U.S today.

FUTURE

We hope to dive deeper into using our data to figure out why some regions such as Italy and New York were hit so hard in just approximately 2 months. We believe that in order to do this, we need much more advanced features such as the age distributions, number of people going out after the shelter in place order, etc.

In the process of figuring out how to answer this question, we are also curious about interesting questions such as: how people can survive after a few months of social distancing (mentally and financially), how to deal with virus rebounds, and how likely the economy will recover after the pandemic.

Other future explorations include feature engineering on the SIR model in order to find the optimal model in predicting the growth of the coronavirus, including factors such as the current shelter-in-place and social distancing procedures.

Other work we also hoped to explore were comparing the reproduction number, also known as the R naught values, of different viruses in the past. Understanding how accurate R naught numbers historically were in mapping the spread of viruses could help us understand the spread of COVID-19.