Project 1: COVID-19

William Mac

6/12/2020

861013944

In this project we are asked to predict the pandemic evolution of COVID-19 based on time-series data which includes confirmed, deaths and recovered globally from Johns Hopkins University Center for Systems Science and Engineering’s (CSSE) GitHub repository. The dataset consists of three files which are confirmed, deaths and recovered. The three files all share “Province/State”, “Country/Region”, ”Lat”, ”Long” and a range of dates as a common header. The only thing different in the three files is that their values. Their values correspond to confirmed, deaths and recovered. As with most cases involving data, the format the dataset comes in will be a problem for us. We will have to address this first before we do any kind of analysis.

When looking at the COVID-19 global time series data we run into immediately noticeable problems. The three files are kept into three different CSV files which makes plotting them in the same visualization difficult. The dates in the files are displayed as a column name, which is a very odd shape to have for a dataset. The values for confirmed, deaths and recovered are not exactly matched because of missing data or the way they reported the data (I.E. some countries don’t report province/state-wise stats for recovered).

Knowing these inherent issues with the dataset, I decided to start out on my python notebook. I decided initially to import the following packages: pandas, numpy, matplotlib, seaborn, plotly and folium. I used pandas to import all three CSV files in their own *respective* dataframe. Then I set a variable called dates equal to a list of the date columns from the *confirmed* dataframe. I used this dates column as the value\_vars in the dataframe.melt function to transpose all those date columns into one column called dates in all three dataframes. After melting, I merged the data together using a left join merge using “'Province/State', 'Country/Region', 'Date', 'Lat', 'Long'” as the join key. I then created a new columns: ‘Total cases’, ‘Total Active’, ‘New cases’, 'New deaths' and 'New recovered'. I also removed Canada due to how they reported their data.

Due to how pandemics can grow out of control exponentially by a factor of 2, we need to identify when a country reaches 100 cases, because it’s only a matter of days before it infects the whole population. we need to create a variable for it which we will later user to analyze how effectively each country handled COVID-19.

Because my data is as of 5/31/2020, I decided to do a gradient map on the combined dataset to get a rough idea of how bad our situation is in the US compared to other nations. Compared to other nations, the United States has almost as much as the other top 8 countries combined in confirmed cases (1.79 million).

I then used folium to map out the numbers for 5/31/2020 on a world map using the latitude and longitude columns. I noticed that Canada was the same size circle all over meaning their data was being reported the same for all of the provinces, so I removed them. I also removed China because their reporting number seem very odd given that they were ground zero, the amount of citizens vs confirmed and how the growth pattern looks in excel (various days look the same which is not possible because of their population, meaning their reporting is less than accurate). I also removed ship data as well since they were not accurate.

I used the same 5/31/2020 dataframe to plot out total cases vs recovered. Seems like the top 10 countries are doing poorly with the exception of Germany. They have a 50/50 ratio. I decided to pick the 5 countries from here and then picked Japan, South Korea and India (since they were all in close proximity to China and they got hit earlier than the US did) and built a dataframe for each one.

With pandemics, the goal is to flatten the curve so that all those infected can be treated by our healthcare system. If the curve goes over the limit of what our healthcare system can provide, that’s when we get people dying who we could have saved due to the lack of resources. From the trend graph, we have a trend of COVID-19 cases in the US. It seems like the growth is somewhat linear meaning it didn’t start off exponential for us and that our curve will not flatten out anytime soon. It will eventually compromise our healthcare system; we just don’t know when yet.

The New Cases graph depicts new cases in the US on a daily basis. Again, this looks pretty linear and falls in line with the confirmed cases.

The total confirmed cases graph compares all the total confirmed cases with the countries I picked. It seems like South Korea, Spain and Japan are about to flatten off. Brazil, Russia and India looks like they are going to keep growing exponentially, hoping to flatten out soon. Italy looks very similar to the US. It could be that because both countries were late to act and were forced to quarantine their whole nation. This is why their graphs look similar while South Korea went to great measures to take initiative in tackling COVID-19.

The cumulative cases graph compares the countries I picked and shows their growth after crossing the 100 case threshold. As you can see, the US does very poorly in this, which is quite sad for a leading country of the world.

I split out the US dataframe into confirmed, recovered and deaths so I can predict each one. The COVID-19 Cases in the US graph shows the number of cases for confirmed, deaths and recovered on a day by day basis globally in a total sum. It’s a reminder for me to not expect any kind of flatting out within the near future. I expect it to continue rising for the next 40 days.

Using all this knowledge from the graphs I plotted, I decided to use Prophet, which is a library that is built specifically to predict using time series data. It’s very robust in that it can easily use MCMC to build an accurate model after setting a few variables such as max\_treedepth to prevent overfitting and underfitting, growth type (linear vs logistic), confidence interval.

I set MCMC\_samples = 300, growth = linear, interval\_width = 0.95 and max\_treedepth = 20. I fit the confirmed dataframe and then loop a cross-validation prophet built in function that does a cross validation for each horizon value (values used to validate) after going over an initial 80 days of data. I then use the output from the cross-validation to get the Mean Average Percentage Error (MAPE) for each horizon value so I can see how the forecasts would look farther into the future.

After seeing that my MAPE for the confirmed dataset was around 12% (88% accuracy) for 10 days into the future and 40%(60%) for 40 days, I was satisfied with that as there can be many irregularities within the data and it makes sense for the accuracy to drop as we predict further into the future. I wanted to confirm my choice by using the performance metrics from Prophet. I decided to plot the MAPE vs the dataframe after cross-validation. From this graph, it looks like we are in the right area to be in.

I also got similar MAPE same for the deaths figure as well. Oddly though, it seems the MAPE is higher for the recovered figures. This leads me to assume that the recovered figures are not accurately represented. On each graph for deaths, confirmed and recovered we are able to predict quite accurately without much deviation except for the recovered graph. That one seems to be the most inaccurate. It could be due to the fact that countries are mis-reporting their actual numbers or that the actual number of recovered is unknown due to people not coming forth about recovery.

Before running the model using Prophet, I predicted that the US will not be flattening out anytime soon because of how the preliminary graphs looked like. It should look like a sigmoid, but it was fairly linear. This led me to believe that the US was nowhere near the end of the pandemic and that the quarantine only delayed the inevitable until a vaccine comes around. As my prediction shows, it is estimated that the confirmed cases will continue to rise gradually with small dips in the middle of the week when most people are at home and predicted to rise around the weekend and after the weekend. This makes perfect sense because people usually go out on the weekend to go grocery shopping and are bound to get COVID-19 eventually.