Covid-19 in the United States

Luke Gruszka

Abstract

Today we are faced with a virus that is rapidly spreading and affecting everyone across the world. While the world governments are taking action and implementing safety procedures; we still see that the virus is spreading. This draws the questions if these actions are preventing the spread or simply just stalling the spread. The data being collected will give me the option to be able to see when the virus will be predicted to end and if the safety measurements being taken are actually slowing down the spread for the virus. By using a Random Forest algorithm, I will be able to map out the trajectory of the virus. Also, I will be able to use a time series to map out when certain safety precautions have been taken and what effects those actions have on the virus outbreak.

Problem Statement

In this study I want to see if the precautions being taken in the United States is really slowing down the rapid pace of the virus. This virus affects the people who fall in the demographic of 70 years of age or older. Some of the issues came from the lag between virus recognition and action taken by governments when first facing the virus. We are unclear when this virus will end and this paper will try to tackle what the next month will look like for the United States.

Background

The coronavirus recently came into light in the world. This virus originated in the city of Wuhan, China and thus has spread across the world. With the recent technology and transportation systems we have seen that the virus has been able to spread faster across 70 different countries. Around December 31st of last year the Chinese authorities alerted the World Health Organization of the outbreak.

The data that I have gathered needs to come with a disclaimer. Tests were not initially available for testing. Tests have driven the conversation and data surrounding the virus. If patients were not tested and could have possibly had the virus without being properly tracked. This is where my model could be inaccurate as far as having the correct data. In the United States the tests have been in a word delayed. The problem in China was not really appreciated until tests were available. The world did not see the true colors of the issue until the tests started reporting back and seeing that the number of cases had spiked.

The data I will be collecting will have to be well labeled and up to date. I looked for a couple different parameters. The only variables that I was curious in was; location, data, cases, deaths. With these variables I will be able to feed my model and make sure that I was answering my problem statement.

Solution

My data was clean and split up into train and testing datasets. This data had to be manipulated a little bit to be able to be as accurate as possible. I transformed my date fields into useable integers because I did not have to deal with string values going into my model. I also dropped the province and state column as I do not need the data to be that locational based as it will not matter for the model. Then I removed the null values from the training dataset as I require all the fields to be filled in. As for the test dataset I kept that as is because it had the data that I was transforming my training set to mimic.

In order to find variable correlation to make sure that cases and deaths were worth pursuing I graphed it out. I took my training data set and graphed the global confirmed cases and the global confirmed deaths.

The only concern that is transparent here is that initially the graph starts out on a straight unaffected line. This could be due to the fact that in the wake of the new virus the healthcare industry was unsure of the true classification of the virus. This in turn will result in false readings and overlooked cases. When I focus my efforts on the United States I will start the data when the virus was recognized as an issue.Some issues that I see that could be affecting the graph would be the availability of the tests for the COVID-19 virus. Looking at the confirmed cases we can notice that the graph exponentially grew around the same time that the tests were distributed to the affected areas. This could affect the model.

I used the scipy library to map out the correlations between the confirmed cases and confirmed deaths. When I used the spearman correlation I received a score of 1, then when I used the kendall tau correlation i got a score of .99. These scores show that there is a correlation between these two variables that when the confirmed cases goes up as does the deaths. But as I mentioned above this could also be reflective of when tests are being administered as that is when they become available in bulk. Or this could be reflective of the lack of response when first faced with the virus.

I want to identify all of the United States data in the datasets to identify a cleaner look. Once the Corona Virus became transparent in the United States there were tests available to test and identify for the virus. I anticipate that this analysis will differ from

the rest of the world, graph above, as the graph will not include China. China's influence in the data showed a delay in progression against the virus when it comes to tests and identification portrayed by the spikes in the data.

When the data comes to the United States I expect for it to start out slow then begin to grow exponentially as more tests will be available towards the beginning of the outbreak. China was unaware of the virus at first and allowed for the virus to spread. As the United States looking in, they should be able to catch the virus early on and take measures early on to prevent a wide spread disease.

Taken from the same timeline we see that there was some time where the virus was not reported in the United States. But once the virus was reported we can see that the cases exploded.

The United States was aware of the virus as it originated in China. On Jan 21 2020 the United States had its first case of the virus. Then a month later Feb 26, 2020 the United States had it first suspected local transmission. Then a short three days later the United States had reported its first COVID-19 related death.

Looking at the graph above we notice that around early March there is a spike in the data. We are able to see that the graph started to grow. This is because on March 3, 2020 the CDC lifts restrictions for virus testing. This means that people will close contact with people diagnosed with COVID-19 or people with severe symptoms could get tested.

This, much like what had happened in China, availability in tests to the general public made it clear that the virus had already spread across the population. Once the testing was presented to the United States and people began to take the tests the volume of the virus came into light. Through the month of march the number of cases grew at an alarming rate because the availability of the tests gave us a clearer picture.

I expect this number to keep on growing as a result that tests are not accessible to everyone. Once there are enough tests to completely test everyone who is concerned with having the virus I expect the data to reflect a steady growth in cases and deaths.

When using the United States data the correlation scores between the confirmed deaths and confirmed cases got a spearman score of .80 and a kendall tau of .76. Even though this is not as strong as the global correlation there is still a correlation between the confirmed number of cases and fatalities in the United States. Even though the numbers

are not similar, they follow a similar pattern and if the number of cases grow so with the number of deaths.

Model

The random forest classification model is an ensemble tree-based learning algorithm. The method is to randomly select a subset of training sets and it will aggregate the votes from different decision trees. This will allow the model to predict the final class of the test object.

The algorithm is stable and will not move or produce significantly different results if a new data set is introduced. The algorithm also does not have any bias. The algorithm can handle missing values and unscaled data points but for this project I have removed those concerns as it should lead to a better result.

Upon using this model I was able to predict the following metrics. By 4/23/2020 the below predictions were made: 1,721,311.0 more people in the US will contract the disease, and as a result I predict that 21,389.0 people will expire. Between these numbers the correlation score was .80 for spearman and .66 for kendall tau. For this score I see a pattern that these numbers could represent valid numbers. Given the timeline with the skyrocketing numbers and the correlations of the predicted number I feel confident in the prediction.

Disclaimer

This project and data had been directly affected by the distribution of tests across the global. In the data and graphs we are able to see the spikes of which the tests became available based on location. In the United States we are able to see when tests are available and when supplies begin to deplete. I also suspect that those spike affected my outcome and the model of which I trained using that data. I suspect that the correlations between the two variables are accurate even if the numbers may be off.

Resources:

1. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html
2. https://ourworldindata.org/coronavirus
3. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
4. https://www.worldometers.info/coronavirus/
5. https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/
6. https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-novel-coronavirus- disease-2019-covid-19-pandemic-increased

7. https://www.barrons.com/articles/latest-coronavirus-data-show-disease-continues-to-spread-even-in- the-u-s-51584224660
8. https://www.theguardian.com/world/2020/mar/13/coronavirus-pandemic-visualising-the-global-crisis
9. https://ourworldindata.org/coronavirus-source-data
10. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
11. https://www.sciencedaily.com/releases/2020/03/200317175442.htm
12. https://www.worldometers.info/coronavirus/coronavirus-age-sex-demographics/