CSC110 Project Report: Investigating Risk Factors for Regional COVID-19 Severity

Angela Zhuo, Deon Chan, Emily Wan, Ian Huang

Tuesday, December 14, 2021

Problem Description and Research Question

As we have seen over the course of the past 22 months, the impact of COVID-19 has affected every nation differently. In larger countries such as the United States, COVID-19 has had a varying impact even on the state and provincial level. In treating a pandemic such as this, it is important that the way that factors specific to every locality affect the transmission, impact, and severity of COVID-19 be well understood.

What are the most important factors that can impact the spread and severity of COVID-19? There have been varying experiences with the rates of COVID-19 which can be estimated to relate to a couple of specific factors. The factors that are most important can be assumed to be geographical location, state, population, population density, GDP, current vaccination rates, and mask usage (Kim, et al.). Geographical location could potentially affect the transmission of COVID-19, depending on regional factors like temperature. Population density is an important factor because people who are in close contact with infected people are more likely to become infected as well. GDP is a possible indicator of how prepared a region is in a pandemic situation. Vaccines have been shown to reduce the chance of becoming significantly ill by more than 90% for people who are fully vaccinated against COVID-19, so regional vaccination rates can be assumed to significantly reduce the number of deaths in an area (CDC). Mask usage is another important factor affecting regional infection rates because when a majority of the population wears masks, this can significantly reduce the rate of COVID-19 transmission (Wang, Yuxin, et al.).

By creating a machine learning model using these factors as inputs, it is possible to analyze how collective factors or individual factors work together to change the condition for the cases and deaths of the coronavirus. This model's predictions can show the infection and death rates of different areas. From these predictions, it will also allow individuals to understand the level of risk of different areas based on changes made to any of these factors.

This research can be helpful for understanding the possible impact another pandemic could have and what factors are most important to keep in mind when estimating this impact. Learning from the past history of the pandemic allows us to make better predictions and models on the number of cases and deaths of future health crises. This is crucial information that can help predict how to limit the severity of future pandemics and how to allocate resources to help areas that are more likely to have severe outbreaks due to hard to control factors. Allocation of proper resources to high-risk areas will alleviate the severity of any possible health crises. This model will allow for these predictions to be made by focusing on specific factors and using past data to create accurate models of infection and death rates.

Dataset Description

We trained our model on a New York Times dataset, containing csv files with US COVID-19 data. This dataset has live and historical data about total cases and deaths per county. The file us_counties.csv contains cases and deaths as integers for each county, recorded periodically at certain dates. We used all columns from this file except for the fips code. We used GDP data provided by the US BEA, in the form of a csv file. This csv file contains the 2020 GDP of each US county in current dollars, in scientific notation with units as a string (eg. 'Thousands of dollars'). From the CAGDP2_ALL_AREAS_2001_2020.csv file, we used the GeoName column and the 2020 column which contains the 2020 GDP of each county in scientific notation. We also used a csv file from the US Census Bureau containing the estimated 2020 population of each county as an integer. From this csv file, co-est2020.csv, we used the columns CTYNAME (county name as a string), STNAME (state name as a string), and POPESTIMATE2020 (2020 population estimate as an integer). We used a 2020 Gazetteer text file from the US Census Bureau, called 2020_Gaz_countries_national.txt, which contains longitude and latitude coordinates for each county. This file also

contains land area in square miles which we used to calculate the population density of each county. From this text file, we used the columns NAME (county name as a string), USPS (state abbreviation as a string), ALAND_SQMI (land area in square miles as a float), INTPTLAT (latitude as a float), and INTPTLONG (longitude as a float). Finally, we used a csv file from the CDC that documents vaccination rate as a percentage for each county. From COVID-19_Vaccinations_in_the_United_States_County.csv, we used the columns Recip_County (county name as a string), Recip_State (state abbreviation as a string), Series_Complete_Pop_Pct (percentage of the population that is fully vaccinated as a float), and Date.

We merged all of the data into a single dataset with each variable put under the single county it is associated with. Then for counties that do not have complete data, we take those counties out of the set to clean up our data. This data is then given to our machine learning model to learn how to predict COVID cases based on the previous data.

Computational Overview

The core of the machine learning aspect of this project is a simple feed-forward neural network that takes, as inputs, geographic, demographic, and economic features represented quantitatively for some arbitrary locality in the United States, as well as a time variable which we will call t. The model, for this set of area-specific factors and some value of t, outputs a prediction for the severity of COVID-19 at that point in time, represented as instantaneous values for the death rate and new case/infection rate in that area. From this information over a fine enough set of values for t, we can draw an estimate of what the COVID-19 graph of total cases and total deaths would look like, not only historically, but also potentially for a small period of time into the future. For this project, our model was trained on a dataset of over 3,200 counties.

The architecture for the neural network we used is quite simple, consisting of a one-dimensional vector "input layer" representing the described input factors, all normalized appropriately to values between 0 and 1, multiple "hidden layers", and an output vector representing our desired instantaneous COVID-19 graph values. The mechanics of a neural network are complex and impossible to comprehensively describe in this proposal; in summary, a neural network is, at its core, a very complex mathematical function with each aforementioned layer represented as a matrix. The "hidden layers" consist of potentially millions of variables that are all adjusted algorithmically using a process called backpropagation in order to best fit the model to our dataset. The size and number of the "hidden layers" are the hyperparameters of the model, and while chosen somewhat arbitrarily, will be tested thoroughly to produce an accurate, yet robust model.

The library we chose to perform the machine learning-related operations in our project is called Tensorflow, an open-source machine learning library developed primarily by Google. The reason we chose this library is that it provides capabilities such as doing computationally expensive matrix operations on the GPU, a task that would be otherwise difficult using ordinary Python code. Another reason we chose this library is that it simplifies the process for the backpropagation step described earlier, a task that would ordinarily be extremely computationally intensive without the use of multivariable calculus.

In terms of data visualization, we allow the user to choose an existing county from a map of the United States as a basis, and adjust the individual geographic, demographic, and economic factors of this area to see how they affect the predicted overall COVID-19 infection/death curve over time for these adjusted factors. We also allow the user to choose how they view the COVID-19 infection/death values, either in terms of their rates of change such as infections per week, or simple cumulative totals in either a linear or logarithmic graph. For this, we used Tkinter, as the library provides more precise programmatic control over how the data is displayed.

Instructions

Before running this program, please make sure you are using a terminal environment that supports colors, otherwise most text will not render properly. Some examples of acceptable terminals include:

a. Any Unix terminal

b. The PyCharm Python console

c. The Visual Studio Code integrated terminal

1. Install all Python libraries under `requirements.txt`.
Please note that Tensorflow may have special requirements depending on your system. For instance, newer Mac

computers running on the M1 Silicon chips may have issues running Tensorflow without installing additional plugins. 2. Run the `main.py` file. Instructions will show up in the Python console. There will be progress updates on the state of download for data sets as well as progress on the machine learning. The output will be an interactive map of the U.S. which will allow the user to select a state and a county and adjust the geographic, demographic, and economic factors. Then, a graph of predicted COVID cases over time is shown based on the inputs that are given. Some of these inputs such as vaccination rate may be manually adjusted to see how their differences may affect predicted COVID impact.

Changes

The biggest change that was made is shifting from using the Pygame library to Tkinter for the data visualization of the project. There was also a change in the machine learning library from PyTorch to TensorFlow due to problems with unstable optimization in PyTorch. We also omitted mask usage data as we did not end up using a suitable dataset with enough data points to train our machine learning model.

Discussion

The graphs that the algorithms produce help show us different predictions for an area based on the variables. Being able to manipulate each variable allows us to distinguish the effects and better understand how each one impacts the spread and severity of COVID-19. From some example graphs, it is easy to see that population density and population both play important roles in determining the spread of the virus. By simply increasing the population numbers, the cases and deaths increase dramatically. That is to be expected because an increase in the number of people in a county also means an increase in who it can be spread to and the reach of the virus is exponentially increased. An increase in the vaccination rate also shows a decline in cases which can demonstrate how it can flatten the curve of cases. The GDP increase also correlates to a higher impact of COVID-19, possibly because a higher GDP can also indicate higher population density.

There are limitations on what we can predict from the amount of data collected. More data can help train the machine learning model to be more accurate with its predictions. Currently, the model's accuracy suffers from overfitting. Additionally, user provided input to the model may produce unexpected results. For instance, we noticed that when raising the GDP of existing counties, the model predicted a significantly steeper COVID case graph. This is an incorrect causation drawn by the model; it is probable that the model associates higher GDPs with denser or more populous areas, and therefore assumes a higher impact of COVID.

Additionally, the predictions that are made are only limited to counties in the US with a full dataset and is not enough to make predictions on how these factors can affect other cities in the world. It is also only possible to create predictions based on changes of a factor but it is unable to predict cases or deaths in the future. We faced obstacles with regards to normalizing the data since the sum case values across the US tend to 0 which made the machine always choose 0 in cases of uncertainty. We had to normalize on a logarithmic scale to be able to have the machine accurately predict the values.

Some of the next steps will be training the model with data from different countries to be able to create predictions for more locations. This would be only possible for countries with similar datasets to the ones collected to create predictions for the US. Another interesting next step will be to predict the case values for those in the future. The last possible next step is being able to predict case values for other similar diseases as COVID-19. This would require the most adjustments as each virus is unique and there are properties that can affect the spread and severity that cannot be predicted by our model.

Through the creation of our model, it has answered the question of what factors are most important and how it is important when it comes to the spread and severity of COVID-19. This model takes in the factors of geographical location, state, population, population density, GDP, and current vaccination rates to help with the prediction. The interactive data visualization helps effectively show the change in impact based on case and death numbers for a single county in the United States. There are some limitations based on the number of data points collected and there are still multiple ways to improve both the model as well as expand the possible predictions. The model is able to do what we had hoped and many of our observations can be applied to understanding the impact of another pandemic in the future.

References

CDC. "COVID-19 Vaccinations in the United States,County." Centers for Disease Control and Prevention, 24 May 2021, https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh.

CDC. "COVID-19 Vaccines Work." Centers for Disease Control and Prevention, 9 Nov. 2021, https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness/work.html.

Kim, H. et al. "Which National Factors Are Most Influential In The Spread Of COVID-19?" PMC, 16 July 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8307075/

Pygame. "Pygame Front Page." Pygame, https://www.pygame.org/docs/.

PyTorch. "About PyTorch." PyTorch, https://pytorch.org/features/.

The New York Times. "Nytimes/Covid-19-Data: An Ongoing Repository of Data on Coronavirus Cases and Deaths in the U.S." GitHub, https://github.com/nytimes/covid-19-data.

U.S. Bureau of Economic Analysis. "GDP by County, Metro, and Other Areas." U.S. Bureau of Economic Analysis (BEA), 9 Dec. 2020, https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas.

U.S. Census Bureau. "County Population Totals: 2010-2020." United States Census Bureau, 8 Oct. 2021, https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-counties-total.html.

US Census Bureau. "Gazetteer Files." United States Census Bureau, 8 Oct. 2021, https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.2020.html.

Wang, Yuxin, et al. "How Effective Is a Mask in Preventing COVID-19 Infection?" National Center for Biotechnology Information, 5 Jan. 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7883189/.