Does staying at home decrease the spread of COVID-19

Predicting the infection rates in US states using mobility measures from the Bureau of Transportation Statistics

Geet Chawla

December 2020

1 Introduction

This projects aims to use machine learning techniques to predict the number of positive coronavirus cases in each state of the United States for a specified time range. By using parameters like population, percent of population staying at home, percent of population not staying at home, number of trips (out of home) made per-capita - the model tries to predict the number of coronavirus cases. The data used is at the daily level from January 2020 until early December 2020. The analyses is trying to determine ”does staying at home decrease the spread of infection at the state level” in the US.

2 Data

Owing to the enormous amount of data collected and maintained relating to the pandemic - this research will be using three rich data sources. Firstly, to accumulate data on mobility - number of trips made at the state level, population staying at home and population not staying at home - we have used the data maintained at The University of Maryland. The Trips by Distance data and number of people staying home and not staying home are estimated for the Bureau of Transportation Statistics by the Maryland Transportation Institute and Centre for Advanced Transportation Technology Laboratory at the University of Maryland.

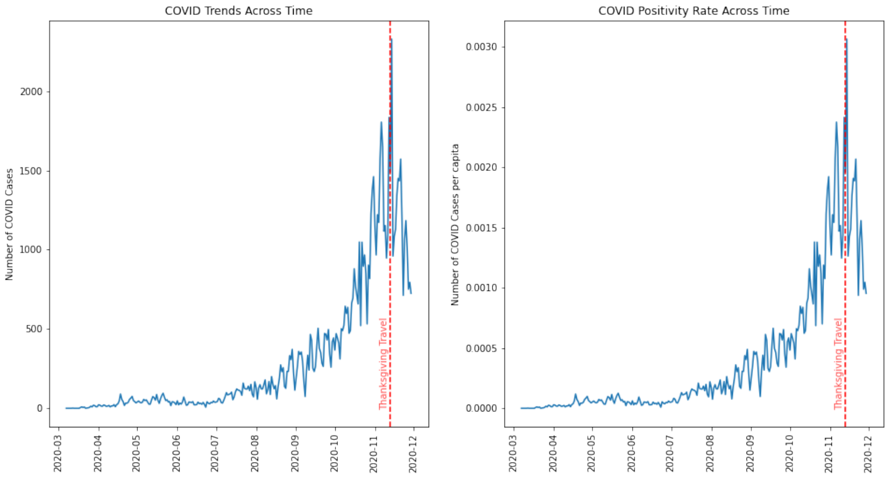
Secondly, we acquired data on the number of coronavirus cases from the datahub repository of COVID databases and dashboards. This data is updated on a fairly periodic basis and presents comparatively latest outcomes. Lastly, to combine and prepare the final dataset, we also get the state FIPS data which we use as a key to merge and manage the above two databases, perform state level analyses, and plot state level plots.

3 Methodology

To determine the correlations we are trying to establish between the number of trips in a state (scale of mobility, as defined by The Bureau of Transportation Statistics) and the number of positive coronavirus cases - we scrape the two data-frames form their original sources using their open source APIs. By directly scraping the data-frames from their original sources - we make sure that this project is updated and the analyses can be run at any point from now. The project will be refreshed as soon as any user refreshes and re-runs the code. This would download the updated data-sets in memory and with the use of further pre-defined functions; any user should be able to regenerate and save updated plots.

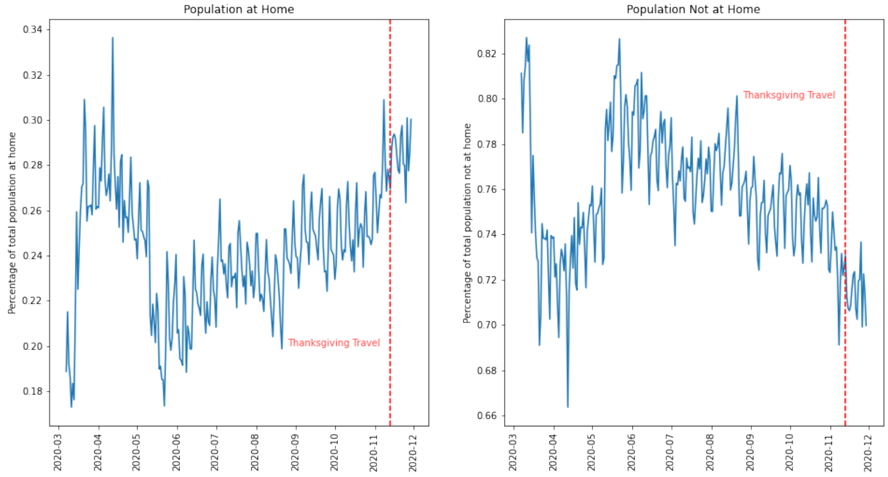
The raw data is then cleaned by performing basic data cleaning steps and making appropriate changes to the data types and structure of the data. Since these two data-frames come from different sources - we add the FIPS codes and State names in both data bases and merge them using FIPS as the common key to create a final database for analyses. The final data has the date, FIPS code, State Name, Population Staying at Home, Population Not Staying At home, Trips taken by the population in a state, and the number of positive coronavirus cases per day - all observations are in the final dataset at the state level. This is a time-series dataset for data from March, 2020 until early December, 2020. These dates in the final data hold true as of 12th December, 2020 - if a user is updating the whole code chunk; they might get even more data.

Figure 1: Time Series Visualizations for North Dakota



From the final dataset, we have created basic visualizations of the time series presenting Covid Trends Across Time, COVID Positivity Rate Across Time, Population at home over time and Population not at home over time. We have also made a special annotation in the visualizations which mark a proxy for the Thanksgiving related travel - this is marked approximately two weeks prior to Thanksgiving. The date in this context could potentially be a good proxy for when people start traveling for the thanksgiving holiday. We have tried to visually analyse the movement in aforementioned parameters around this time and we can see obvious trends of increase in the spread of the infection around that time. The data in this summary are presented for the state of North Dakota. Using the code and predefined functions; such visualizations can be made for any state of the United States for the latest available timeline.

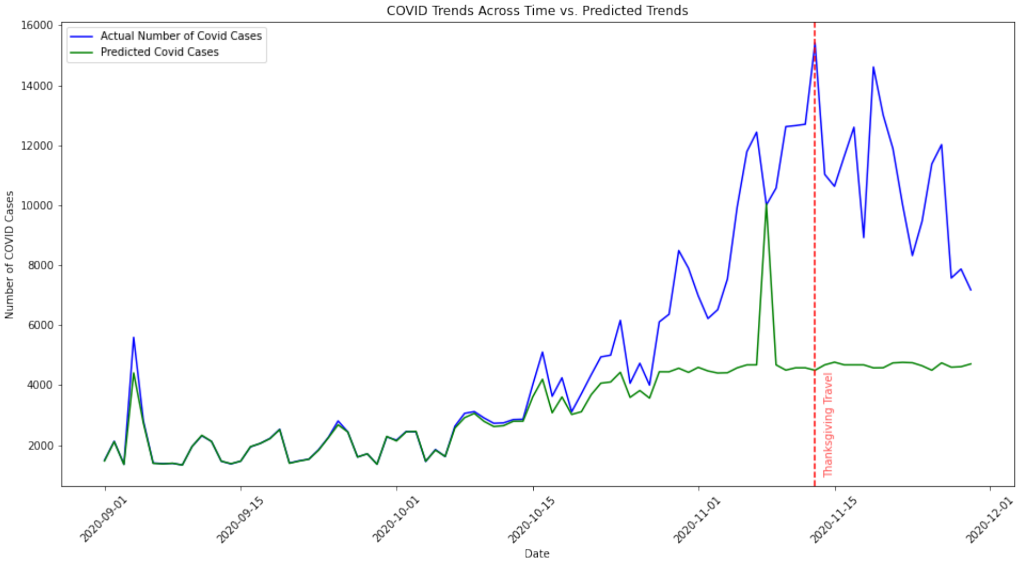
Figure 2: Time Series Visualizations for North Dakota



We then move ahead to use a machine learning model to perform predictive analysis for the COVID cases. In this analysis, we fit a model using population at home, population not at home, and number of trips as the features to determine the positive number of COVID cases in a state. Since the data is very large in magnitude and scale, we decide to skip the testing of models like the OLS, SVM, and Decision Trees as it was becoming computationally very heavy to compute the matrix inverses and other mathematical functions in Python. We still rather decided to chose a fairly complex model of Random Forest. In this method, the algorithm iterates over the features and its connections in the forward and reverse time periods a substantial number of times to define the probability of the next step (or, the weights). In this case, with those weights, we then forecast / predict the future values / number of positive coronavirus cases.

To train our model, we have used data from March 2020 until August 2020. The validation data starts from September 2020 until early December 2020 (as of 12th December, 2020). We then go ahead and create a pre-defined function to attach the predictions of the COVID cases from our model right next to the actual (or observed) COVID cases in order for us to be able to visually compare the performance of our model.

Figure 3: Time Series Predictions for Illinois



The above figure shows a sample of our results for the state of Illinois with actual cases in blue and predicted cases from our model in green. This also has a red annotation for the Thanksgiving travel proxy - which would have helped us understand the impact of holiday travel on the spread of the infection.

4 Results

The results and analyses in our research can be interpreted in two ways. If we only visually try to make sense of the predictions across the actual cases, it may suggest that our predictor is actually not doing that great of a job. The green line seems to be way off than the blue line in Figure 3 (and other samples that we can see using other FIPS in the code). Similarly, if we see the proxy for thanksgiving proxy, we do not see a clear causation because of travel as the cases started rising almost a month before the thanksgiving travel proxy.

However, if we see the predictor from September 2020 until mid-October 2020 - we observe that the predictor is doing a great job at predicting the number of COVID infections in that period. This makes us think that the second wave of coronavirus (which is the hump towards the end) is potentially not just caused by the thanksgiving travel and mobility of people - there could be other factors behind the second wave which can explain the later half of Figure 3 more accurately.

5 Conclusion

We saw that predicting COVID cases only by accounting for mobility as a measure may not be the best features (for all of the time period). This pushes the scope of this project and encourages the use of other features like mask mandate policies, adherence to social distancing norms, outdoor and indoor dining norms, among other precautionary measures. Adding these features in our feature matrix predicting COVID cases might improve the predictions, especially during the second wave of the infection in the US.