Projecting COVID Cases Anywhere in the US on Laptop Grade Hardware

Presented by exi

Introduction

As the COVID-19 pandemic rages, thousands have started working on various methods to forecast the spread of the virus. While leveraging the scope and depth of the C3.ai COVID-19 Data Lake, we guided our approach with basics about epidemiology. Our model focuses on case count rather than deaths due to the financial, psychological, and economic burden that comes with a COVID-19 diagnosis. In the end, we developed a location and population agnostic model that can be trained and deployed for any region. Our model is accessible, accurate, and easy to train.

Problem Description

In our initial research, we found that most efforts in using AI to model the spread of COVID-19 were computationally expensive. Furthermore, data regarding the spread of the virus was in disparate locations and it was a challenge to bring it all together to create models that could inform policy and decision making. According to an online leaderboard, the best models have only beat a baseline estimate 92% of the weeks with a single week percent error of ~2% at best - these aren't even the same model. And with the C3.ai COVID-19 Data Lake, we have developed a model that is better. Our model is further enhanced by looking at data down to the county level rather than the state or national view. The level of granularity we are looking at can be viewed in the aggregate to gain a more robust view of the pandemic and its impact at any location scale.

Broad Approach

We created a model using mobility data, case counts, death counts, and compliance with various public health and safety measures, to predict the number of new cases of Covid-19 per day on a county-by-county basis.

The mobility data came from Apple, Google and Place IQ through the C3.ai COVID-19 Data Lake. Case counts and death counts along with population data on the county level were also used as inputs for the model. The mobility data shows how much people are moving around in a given region and we considered that it could be a good predictor of future case counts as we were essentially looking at how many different vectors the virus would be able to propagate through. Looking at survey data concerning people's likelihood of adhering to social distancing protocols, mask wearing protocols and handwashing gives us an analogue for infectivity to add another dimension to our model. By looking at how people are moving around and their willingness to adhere to viral transmission procedures, we can model how the virus will spread in a region based on local data.

Technical Details of Approach

We are using a highly targeted approach with mobility data and compliance with social distancing measures survey data. We deploy a Gradient Boosting Matrix (GBM) model. With the GBM model, we were able to rapidly prototype a model that could be easily trained on a laptop grade CPU. Our results indicate we achieved a ~2% error on a single county prediction of cases 4 weeks in the future (days 23-28 where day 0 is today) with multiple trials (LA county. Top 5 most populous counties vary 0-7%) Our model outputs a scalar value which is the average of the new cases in days 23-28 reported as a percent of county population. In our demo, the number you see is a rough approximation of percent error (mean absolute error/mean absolute correct target). From here we can go to raw case numbers, death counts via current death rate and see cumulative values with simple math.

Results

For the most populous county in the country, LA county, we achieved an accuracy of ~2% error over multiple trials. Other counties are in a similar ballpark. Counties where survey data is unreliable see errors from 5-10%.

Impact

Currently, the best models as seen on a github leaderboard populated with CDC models only achieve better accuracy than a baseline model 92% of weeks (national predictions). Other models achieve a max of 2% error in a single week with error getting as large as 10-15%. We achieve superior accuracy with 2% error for LA county, and <10% error for the top 5 most populous counties in the US. We present a flexible executable program which can be leveraged by anyone anywhere with very limited hardware. Adding county data together nets national predictions. Our GBM model is not only more accurate than leading models in terms of percent error, but achieves its accuracy with a lower computational cost compared to deep learning frameworks. Our model can be trained on computers of various speeds from the last decade. We hope that public health departments across the country will adopt our model after validating its results internally.

Our model predicts at a more granular level and more accurately than current national and state level models. With predictions nearly a month in the future and this accurate, public health officials can take decisive action before cases explode - saving lives, untold suffering of survivors, as well as millions, if not billions in economic damages. Our model could be the difference between a family getting torn apart or a child becoming homeless. Preventing a pandemic at the local level can help avoid the social and economic fall out far before national trends have even changed. Now, local leaders are armed with the latest in data science thanks to c3.ai’s unparalleled data lake as well as our innovative and lean model.

Here is a link to the leaderboard on github. The models presented here send predictions to the CDC and they are evaluated.
https://github.com/youyanggu/covid19-forecast-hub-evaluation
They are submitted here:
https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html