DSGA-1003 Final Report: COVID-19 Analysis

Noah Kasmanoff (nsk367) , Nathan Griffin (nlg297) , Peter Simone (ps4021), Rahul Zalkikar (rz1567)

Introduction

The COVID-19 pandemic is an unprecedented global event. It has tested the leadership and preparedness of countries and their leaders across the world. In the United States, the vast amount of available data can allow us to shed some light on possible issues as well as possible solutions. Throughout the outbreak and the response, various state-wide mandates have come into effect and we have observed how different regions of the United States have reacted to the ongoing pandemic. Additionally, it has become clear how scarce some resources are, which are crucial in combating the spread of the virus. We perform exploratory data analysis to gather insight related to the COVID-19 response efforts, which will help shed light on some critical components that impact the spread of the virus through different regions. We also aim to create a risk index that can help in the allocation of resources intended to fight g the pandemic. Understanding what leads to the spread of the virus, and understanding how we can properly distribute necessary resources such as tests and medical supplies are valuable components of this paper.

Data

Pandemic

The most vital data for this project is daily new cases and new deaths at a county level. John's Hopkins[2], has updated this information daily and provided this data in the form of a GitHub repository.

Another important trend to take into account during the pandemic is how foot traffic has changed. This data is obtained daily from Google's COVID-19 Community Mobility reports [3]. This data contains information at the county level of how foot traffic at various public spaces, from retail centers to parks, has fluctuated from baseline values before the pandemic.

We explore how policy intervention affects the spread of COVID-19. Worldometer [9] allows us to monitor the dates and descriptions of how different state governments took measures to control the outbreak.

Additionally, we examine how one of America's most vulnerable groups, prisoners, are influenced. A topic of much discussion in the news at the beginning of this pandemic was whether or not to release prisoners with low-level offenses from jail, so as to not risk the unnecessary spread of to disease. We utilize [4] to track the daily populations of prisoner's as another factor in this disease's spread.

Demographic

There is valuable information within static data as well, which helps define a given county or state. To gather population, demographic, and socioeconomic data, we utilize the US Census API. [1] Wikipedia was scraped to find county areas, locations (latitude/longitude), and FIPS IDs, which are used as unique identifiers for each county. [8]. This data is quite straightforward, and is static for each date throughout the dataset.

Cleaning and Engineering

All data, after being scraped, was cleaned accordingly and merged based on FIPS and Date. Many values were missing in some of the pandemic data, particularly within the Google data, since it started being gathered at a date much later than when COVID-19 cases were first reported. Additionally, there would not be enough data for Google to report results in certain instances. Lastly, there is about a 4 day lag on the google data as well, so data is missing for the 4 most recent days [3]. All missing values were interpolated and extrapolated, by county, when possible, and by state, when necessary. This work resulted in a final merged dataset incorporating each data source with no missing values, and therefore analysis and modeling could be performed.

Considerable effort went into engineering features. Since we model all counties, we normalize statistics to be a percent of population size, when possible. Furthermore, we normalize positive cases and deaths to population size. This has the effect of normalizing all the data to the same scale. A proximity metric was engineered as well by determining relative distances to "big cities" based off latitude/longitude. Importantly, the target variable of aggregate positive cases was normalized by population size, which transformed it nearly to a normal distribution. This makes it feasible to eliminate the need to transform values from the prediction, or to predict values that have not yet been seen in the data.

Approach

Exploratory Data Analysis

Our initial goal with EDA is to confirm or refute some of the snippets that have been reported frequently in the national media. For example, that the black population is overly affected compared to the white population. Beyond just looking at the impact, we can see if there are any underlying factors that may contribute like earnings, travel time to work, etc.

Another topic of discussion throughout this crisis that we are interested in exploring is the response of various state and local governments, and the effectiveness for strict measures being taken to prevent the spread of COVID-19. Specifically, we compare the responses of New York and Florida, and see how effective two seemingly opposite policy moves are. New York took swift and strict action, while Florida never issued a strict stay-at-home order. Examining pandemic data, we can compare these decisions.

Risk Model

As mentioned previously, resources critical to fighting COVID-19 are scarce. Testing is scarce, but becoming more available. An objective is to effectively distribute supplies to areas at higher risk. The idea of risk is slightly confounding given population sizes, locations, etc. We define risk as the spread of COVID-19 with respect to population size. For example a 2% increase in a highly populated county may result in more new cases than 50% in a small county, but the 50% would be considered much more at risk.

While it may seem natural to frame this as a time series problem, we frame it as a supervised learning task. By lagging the time-dependent features, we use information from previous days as features in predicting new ones. We incorporate the previous 3 days as features in the model, in an attempt to predict a future event. Further, there is much higher utility in predicting outcomes 7 days in advance, as opposed to 1 day in advance, given the amount of time it would take to reallocate resources based off these results. The idea is that, given a snapshot in time, we are able to identify higher risk counties a week in advance, where a county's risk is a relative metric, comparing it to all others

The target feature was engineered as the aggregate positive count of COVID-19 cases for a specific county, normalized to it's population size, and then shifted to capture it's corresponding value 7 days ahead. This same metric was lagged 1, 2, and 3 days to be used as features in the model as well. `AdaBoost` was used as the optimal model type, with a decision tree as a base estimator.`AdaBoost` is an additive gradient boosting algorithm, where the optimization occurs in the function space and sequential base estimators are fit on the residuals of the previous base estimator.

After structuring the data appropriately for a supervised learning approach, we'll next discuss the cross-validation method. Rather than a standard k-fold cross validation, we use a nested time-dependent cross validation method, which is shown to be better suited for supervised models where there is a clear dependence on time [6]. We subset the data on a window of dates, train the model on this subset, and see how well it performs on a small series of dates following this window of time [6]. Then, iteratively, the timeframe window is shifted until all points in the training/validation set have been reached. The idea is that, assuming heavy time dependence, we could train a model at any window of time in our data, on selected hyper-parameters, that would be able to generalize well to future events. Each split of the cross-validation process produces an R^2 value based off a small set of dates following the window of dates trained on, and we'd ideally select a model where this metric is optimal across each split. If we can find a model that generalizes well, regardless of which stage in the progression of COVID-19 we currently are, then we would select that model.

The hyper-parameters tuned for the base estimator decision tree regressor is just `max_depth`, which controls how deep a tree can grow. The hyper-parameters tuned for the `AdaBoost` model are `n_estimators`, loss, and learning rate. `n_estimators` controls for how many additive base estimators are fit, while loss controls the type of loss (linear vs. square), and learning rate just controls how much impact each model has on the combined additive model. A slight caveat to this method is that, given the nature of the COVID-19 data, earlier time points don't have nearly as much data in them, while recent dates have far more data, so cross-validation splits on earlier dates are likely far less reliable than recent ones. We evaluate 10 "windows" of time in our cross-validation, which, in their entirety, span the full range of dates in the training/validation data, but pay closer attention to those more recent windows of time, given the amount of data at each stage. A brief breakdown of the cross-validation is present in the Appendix in Table 1.

Experiments

Exploratory Data Analysis

Our first piece of EDA involves looking at correlations between total positive cases normalized for population and some variable of interest on the most recent day in our dataset. In other words, taking the approach of looking

how counties have fared up to this most recent day. From 3, we see the correlation between the white and black population with cases, and confirm that the populations with a higher percentage of black Americans are hit harder by the pandemic. Another highly correlated variable is the use of public transit, which would confirm the intuitive hypothesis that people gathering in one place exacerbates the spread of the virus. Also related to commuting behavior, the longer people have to travel to get to work, the worse the outcome for that county; though length of commute is not as correlated as the mode of transportation.

Delving further into the outcomes of the white vs black populations, there are a few interesting things to note. First, counties with a higher proportion of black population are highly correlated, relative to other features, with poverty rate and travel time to work. One of the highest negatively correlated features, from all features in our dataset, is percent in workforce. With all this information in mind, it seems the black population is being overly affected due to a lack of economic stability. It is harder to stop working when already in poverty, or when available jobs in your county are scarce as is. Further, by having to travel longer on average to get to work, there will be exposure to a wider range of people, which can speed up the spread.

As mentioned, the value in public decrees has been brought under fire in many states. Two states in particular, New York and Florida, have seen considerable publicity for taking two contrasting approaches. As such, we can use the collected data to evaluate how such decisions turned out. Using Google's COVID-19 mobility report data [3] and JHU's COVID-19 data repository [2], we observe what influence these mandates had on foot traffic and how these policies directly affected the relative number of reported cases and deaths. The identified relationships are highlighted in figures 1 and 2.

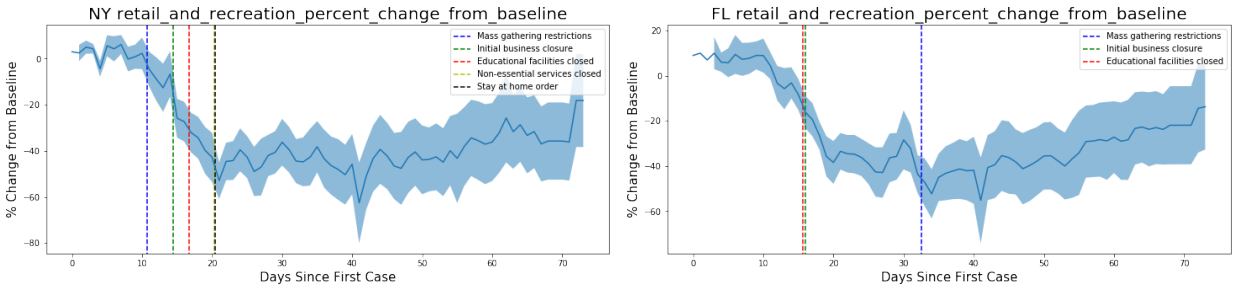

Figure 1: Change in retail foot traffic from baseline prior to COVID-19 outbreak. Noting the times of the mandates in place

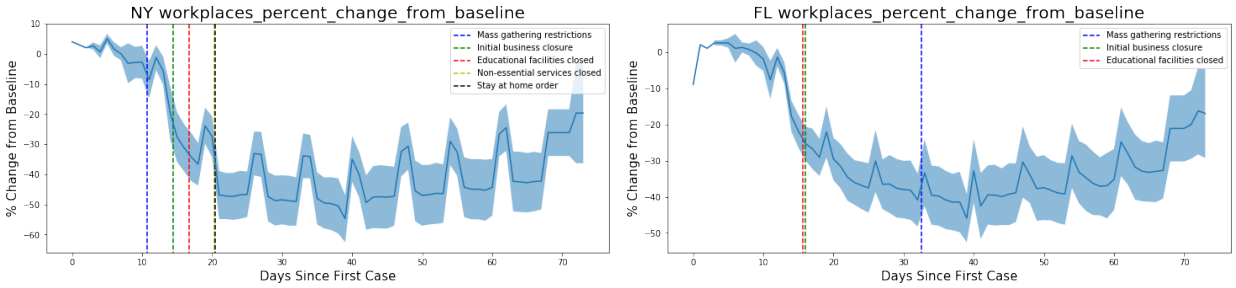

Figure 2: Change in workplace foot traffic from baseline prior to COVID-19 outbreak

Stay at home orders, mass gathering restrictions, and other mandates are highly politicized. [7] [5]. We briefly explored the trend in new positive case growth after policies are enacted in highly impacted counties and did not find significant evidence that a stay at home order caused a change in the growth rate of new positive cases normalized by population per day. In addition, we see in 1 and 2 that despite New York issuing a stay at home order over 10 days before Florida issued mass gathering restrictions there was a similar downtrend in retail and workplace foot traffic in both states.

We conclude that other continuous data, such as the percent change in retail foot traffic might explain the same variance in our target variable as the enforcement of policies.

Risk Model

As mentioned, we use a time-nested cross-validation method testing several hyper-parameters for both the base estimator decision tree regressor and the `AdaBoost` model itself. Results for cross-validation using the finalized hyper-parameters are shown in the Appendix in Table 1, where we include the amount of training/validation in each window as well.

The dataset is split into a training/validation set and a holdout set. The holdout set consists of the 15 most recent days in the data, and is used to evaluate the out-of-sample performance on the data of the final model. Cross-validation was performed on the training/validation subset as mentioned above, and the following hyper-parameters were considered optimal: `max_depth` = None (sci-kit learn default, allows the tree to grow until all nodes are pure, or number of samples in each leaf is < 2.) ; loss = 'square' loss ; `n_estimators` = 10 ; `learning_rate` = 0.1. Finally, the training/validation data was combined to train the final model, using the optimal hyper-parameters found during cross-validation. The out-of-sample R^2 value was 0.9266, which again, is based off the out-of-sample most recent 15 days. This seems rather high, but a 15 day holdout set is quite generous, especially given how much data exists at the most recent dates. An example of the output of the model for May 14th, 2020 is presented in Figure 3, where results are normalized to indicate the scale of relative risk. Given that we predict relative risk 7 days in the future, Figure 3 visualizes the relative risk of May 21st, 2020. This can be valuable in trying to understand how to reallocate resources across the United States.

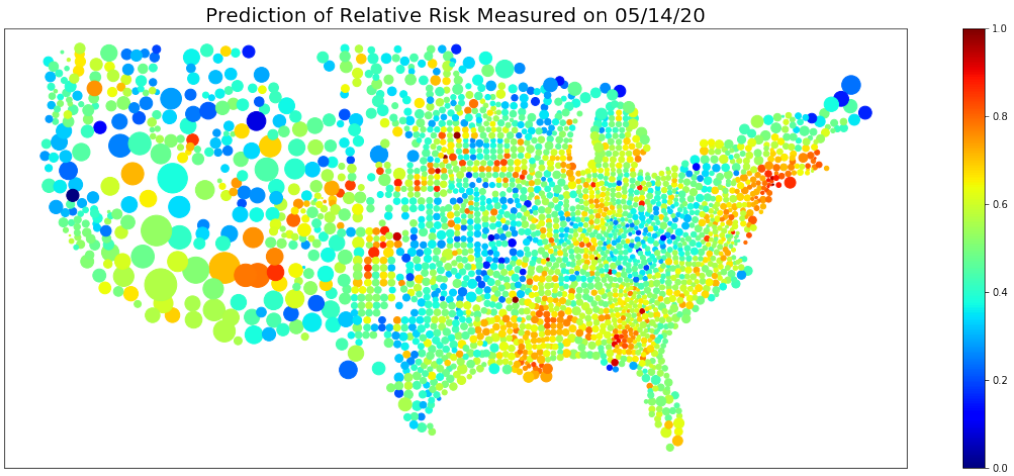

Figure 3: Predicted outcome measured on May 14th, 2020. Since our predicted values are for 7 days ahead, this is what the predicted relative risk looks like for May 21st, 2020. Each point represents a county, the size of the point represents the square mileage of the county, and the color represents the relative predicted risk

Conclusion

The `AdaBoost` model, with non-parametric decision tree regressor base estimators, is able to predict the number of positive COVID-19 cases normalized to a county's population size quite well. This normalization of the target feature was important in reducing the need for predicting unseen values. We were able to successfully model this as a supervised learning task, and it shows the ability to generalize well enough to predict upcoming days successfully. This model could be useful in redistributing resources to those areas at higher risk. The features that the model deems most important are outlined in the Appendix in Table 3. To no surpised, previous day's count is the most important by far. The google trend data is of relative high importance as well. What is interesting is that the encoded variables for time of statewide mandates coming into effect are of little to no importance in the model. An explanation of this is, if looking at Figures 1, 2, the downward trends in foot traffic are already taking place before most mandates come into effect. Also, some states don't even have certain mandates in place, and their foot traffic trends align very closely with those states who do have mandates. Essentially, these mandates do not influence the data, since the data shows that people are generally taking their own precautions regardless. All of the social distancing and foot traffic trends that we observe through the google data already capture the effects that we'd expect to see encoded by time of statewide mandates.

A caveat to mention up front is that number of positive cases reported is a function of testing capacity, so the curve that is followed in our data may not actually replicate the actual spread of the virus. In the future, a more clear picture may be available.

The model and exploratory data analysis presented in this paper can be used to help inform preparations for the next pandemic of a similar nature. Namely, to deliver policies which may reduce the burden on those who are not financially stable and to preemptively model case growth. Further, it seems as though the state-by-state mandate approach was not practical or efficient, so centralizing dialogue and best practices for the whole country may be a better option.

Appendix: Data

Time Window Train	Time Window Val	Training Size	Validation Size	Validation R^2
01/29/20-02/29/20	03/01/20-03/04/20	227	44	0.5956
01/31/20-03/06/20	03/07/20-03/10/20	291	74	0.6287
02/06/20-03/12/20	03/13/20-03/16/20	410	325	0.6756
02/12/20-03/18/20	03/19/20-03/22/20	1028	1291	0.7062
02/18/20-03/24/20	03/25/20-03/28/20	3269	3258	0.6676
02/24/20-03/31/20	04/01/20-04/04/20	10239	6690	0.7355
03/02/20-04/06/20	04/07/20-04/10/20	20816	8851	0.7862
03/08/20-04/12/20	04/13/20-04/16/20	34352	9989	0.8416
03/14/20-04/18/20	04/19/20-04/22/20	49328	10564	0.8954
03/20/20-04/25/20	04/26/20-04/29/20	67105	10896	0.9293

Table 1: Time nested cross-validation, with the corresponding time windows, training/validation sizes, and validation performance at each of the 10 splits

FEATURE	CORRELATION
DRIVE ALONE TO WORK	−.203
CARPOOL TO WORK	.037
PUBLIC TRANSIT	.334
BUS	.165
WALK	.014
OTHER TRANSPORTATION	.030
<15 MIN COMMUTE	−.083
15 TO 45 MIN COMMUTE	.051
>45 MIN COMMUTE	.074
% BLACK	.192
% WHITE	−.220
% OTHER RACE	−.088

Table 2: Feature correlations to total positive cases.

FEATURE	IMPORTANCE
RESIDENTIAL PERCENT CHANGE FROM BASELINE	0.0031
POPULATION	0.0036
POPULATION DENSITY	0.0039
GROCERY AND PHARMACY PERCENT CHANGE FROM BASELINE	0.0047
WORKPLACES PERCENT CHANGE FROM BASELINE	0.0054
RETAIL AND RECREATION PERCENT CHANGE FROM BASELINE	0.0061
POSITIVE CASES POPNORMED SCALED LAGGED 3	0.0104
POSITIVE CASES POPNORMED SCALED LAGGED 1	0.8634

Table 3: Feature importance of risk model

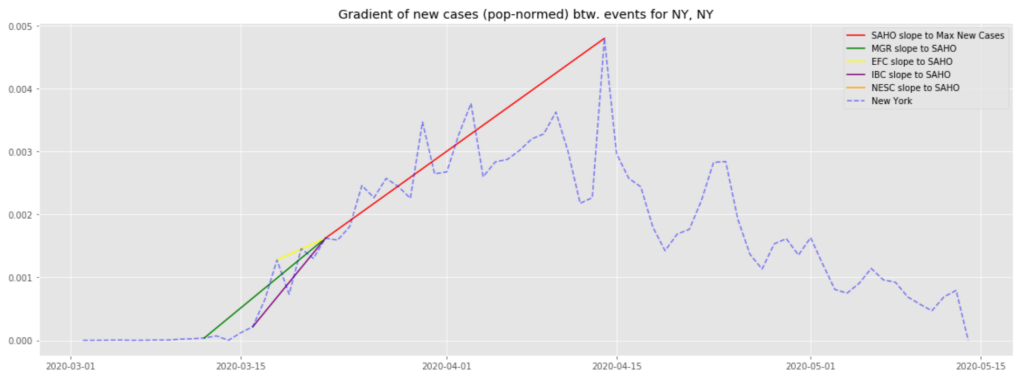

Figure 4: Gradients of Normalized New Positive Cases over Days across Events. SAHO: Stay at home order, MGR: Mass gathering restrictions, EFC: Educational facilities closed, IBC: Initial business closed, NESC: Non-essential services closed (same time as SAHO)

References

[1] United State Census Bureau. Developers.

[2] CSSEGISandData. Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university.

[3] Google. Covid-19 community mobility reports.

[4] Vera Institue of Justice. Vera researchers.

[5] University of Kentucky and University of Louisville. Stay-at-home orders slowed covid-19's spread in the us.

[6] G. Athanasopoulos R. J. Hyndman. Time series cross-validation.

[7] UW-Madison. Study finds stay-at-home order is flattening curve in wisconsin.

[8] Wikipedia. County (united states).

[9] Worldometer. Covid-19 coronavirus pandemic.