Project 3 - CSCI 5751

Trevor Winger

May 6, 2022

1 Abstract

Managing public health is a challenging problem; besides quantifying health risks, public health officials must navigate highly complex political, social, and economic landscapes. During the COVID-19 pandemic, the federal government has offered broad guidance on public health but has left the primary public health work to state and local governments. Deferring public health management to more micro-levels has resulted in significant differentiation in pandemic severity across the country. In this project, I hope to explore phenomena pertinent to showing discrepancies between pandemic management and apply methods in machine learning to observe commonalities and patterns in pandemic management.

2 Introduction

The coronavirus pandemic has been challenging on many fronts. As a result, institutions and existing social infrastructures have been pushed to limits not observed in contemporary public health. Stimulus efforts were given at the federal government level in direct payments to citizens and loan programs to protect private infrastructure. Medical infrastructure had several requirements that stress-tested the current system; this included separating COVID-19 patients from general patients, increased requirements in protective equipment, and less capacity due to spacing requirements (many institutions kept only one patient per room). Existing institutions also had to change their normative modes of operation. Colleges and universities had primarily shifted to online methods of administrating courses. Non-essential employees were required to work from home when possible in certain states. Certain institutions, such as prisons and correctional facilities, could not transition to remote options. As a result, there were many severe outbreaks among the population (both staff and general populations within the institution).[1]

While work has been done on evidence-based management related to the COVID-19 pandemic, little analysis has been done comparing longitudinal data across micro-levels to analyze distinct states' abilities to manage a public health crisis. [2] By utilizing data recorded at the state, county, college, and correctional facilities levels, we believe that we can make definitive conclusions about people's behavior during the pandemic and the effectiveness of public health measures taken at the state level. These conclusions will be drawn by analyzing the percentage of cases from each subsector recorded and creating a pandemic profile for each state by looking at daily volatility across deaths and infections. We also look to utilize methods in machine learning to observe features that may contribute to worse pandemic conditions than other public health management features.

3 Methodologies

3.1 Data Loading & Preparation

All data used in this project is publicly available at the New York Times' GitHub repository. The data was uploaded to our s3 buckets and made public. At the same time, it is possible to run git commands and pull data from within Databricks. To ensure stability throughout the submission process and across workspaces, we opted for the safer option of managing storage and credentials ourselves. The schema is inferred by Databricks (verified correctness by looking at the suggested type), and the header is parsed by built-in spark functionality. A substantial amount of preparation included renaming columns not to be ambiguous on joining data frames for more robust processing, i.e., 'cases' on the college and state data frames being renamed to 'college_cases' and 'state_case', respectively.

3.2 Validation & Aggregation

Given the longitudinal nature of the data, a decent amount of processing had to be done to make data able to be consumed by a classifier. Validation was done by looking at statistics at each state's county, college, and correctional facility level. The system looks at summary statistics for each partition of data. It ensures no significant outliers, i.e., a college being responsible for 20% of cases for an entire state. Overall, the data was curated well, and minimal errors were found. An example that the system discovered was in Minnesota; a day for Cook County accounted for 13% of the cases in the state, while the county occupies less than 1% of the state's entire population. Based on the percentage of cases, this filtering process may exclude some early data where cases are concentrated in a primary location. Still,

we felt that this was an decent decision because we are building a profile for the entire pandemic. The majority of the pandemic management does not occur in the initial infections segment.

To make this longitudinal data consumable for predicting pandemic management, a significant amount of statistical processing on the time series data needed to be done. Given that the dataset has features available daily, creating summary statistics for each day seemed to be a good point for analysis. This included each day available, computing the percentage of cases from each county, college, and correctional facility. In the cases of the county and correctional facilities, rates of deaths were taken as well. In the case of the prisons, we were also able to compute the percentage of inmates that die once contracting COVID. Once summary statistics have been calculated on the dataset, we can build a profile for each state based on these. For each feature computed, we take the average and standard deviation of the percentage of cases and rate of deaths and join these summary statistics with the sentiment recorded about masking across the state.

The difference between deaths each day divided by the difference between cases in the state was computed for labeling, and the average was taken. Measuring the day-to-day volatility was the goal of this assignment. This metric seemed to reflect best the extreme nature of managing public health in a pandemic environment. Turning this problem into a binary classification problem, we split the data based on some threshold of this daily death volatility marker. The thresholds used were: average, average + (standard deviation / 2), average - (standard deviation / 2). The average would be a perfect partition of our dataset into two; the other would split are data into approximately a 60-40 percent split each way.

3.3 Pandemic Profile Justification

The feature set computed to build up pandemic profiles for each state includes enough features that we believe can make our models more agnostic to population bias than other attempts. For example, in Figures 1, 2, and 3, we can observe some interesting trends in average college cases in states that have been labeled as not managing pandemic spread well. For example, Texas, a state with a high amount of disease and death, has had a limited amount of transmission in a college setting and has had a high sentiment toward masking. In contrast, North Carolina has a significantly higher infection rate at colleges while sharing similar feelings towards masking and substantially less variability in day-to-day transmission and deaths. This observation can give us a little insight into the difference in behavior

in each state. It is reasonable to think of drawing multiple conclusions here, such as: perhaps the university population is more extensive in North Carolina or that the University system may have a much more integral part of North Carolina's social fabric and may be more integrated with society as a whole. Detecting nuances like this has powerful implications when predicting management, and moving forward could offer guides at a state level for better restriction suggestions based on the trends within the state rather than general overarching guidelines.

4 Results

4.1 Data Insights

Interesting insights can be derived from correlation analysis of the computed feature set. In particular, the label mechanism was generated as a function of death and transmission from day to day. Still, these features seemed to have a statistically significantly lower when compared to features pertinent to things like masking sentiment. As seen in Figure 4, the sentiment features (averages and standard deviations all have higher levels of correlation with the label than any death or case feature). Another telling insight is that the standard deviation across all sentiment features has a higher correlation than average. This suggests more variation in sentiment towards precautions pertaining to more volatile pandemic conditions. This seems intuitive; with higher variations across sentiments in the rare and never sentiment classes, things like large gatherings with improper protection may be more prevalent when compared to states where these features have fewer fluctuations. After sentiment, the most statistically significant features include the average cases and death at the county level. This observation is also relatively intuitive. County behavior is probably much more reflective of macro trends than college and correction facility levels.

4.2 Machine Learning

4.2.1 Model Description

As it pertains to model development, we could make a couple of assumptions before development: the first being there is insufficient data to utilize methods in deep learning, given the similarity throughout the dataset that an unsupervised learning model may be able to gain some interesting insights. With this in mind, we decided to utilize a stable of models and compare their performance across standard metrics such as accuracy, precision, recall, and f1 score. The models chosen were: a nearest neighbor algorithm, linear support vector

machine, radial basis function kernel support vector machine, Gaussian process classifier, decision tree, random forest, a neural network, Ada boost classifier, Gaussian Naive Bayes classifier, and a quadratic discriminant analysis classifier. In addition, some measure to prevent overfitting was done, like setting a max depth of 5 for the decision trees and max iterations for the neural network of 500.

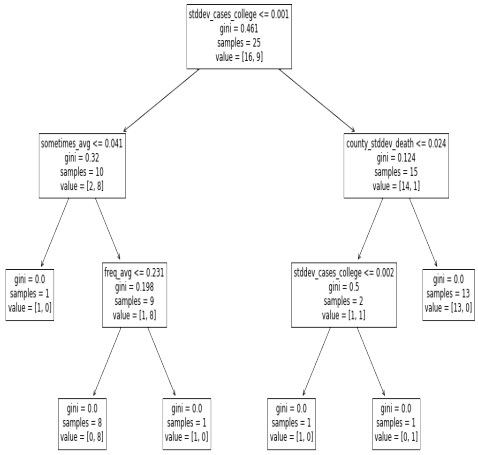
4.2.2 Model Performance & Observations

Across the three label splits, the most exciting results occur in the partition at the mean. Models overfit when the class is imbalanced due to the limited size of the data set. We suggest synthetic data set creation or aggregated data at a similar granularity level worldwide.[3] The top three classifiers and their performance can be found in the table below.

Classifier	Accuracy	Precision	Recall	F1
Ada Boost	.84	1	.5	.67
Decision Tree	.76	.58	.87	.7
Random Forest	.72	.6	.375	.46

Ada Boost probably performed best, probably in part due to its ability to learn quickly in a small dataset; the skelarn implementation utilizes decision trees as the weakest classifier under the hood for the parent algorithm to learn from. The results seemed to adhere to our original hypothesis that the dataset is too shallow for deeper learning methods. The rule-based generation learning algorithms would perform better.

Decision trees offer insights into feature significance by displaying the features used to split the branches. This type of analysis may reveal trends not observed in normative statistical analysis but instead allow observations on the feature set's entropy related to our label class. For example, looking at the tree below for the mean split data, we can see that the first feature split is the variation in cases at the college level, followed by the sometimes sentiment average and the county death rate standard deviation. If we were to guess based on the Pearson correlations, the intuition is that the splitting factors would be primarily on the sentiment feature set. Still, those feature sets have less predictive power than other daily summary statistics. It is unsurprising to see classifiers utilize standard deviation more than mean; we believe this is a much better indicator of pandemic variability when compared to averages. Standard deviation is also less skewed by population and maybe a better predictor of the public



health measure.

5 Conclusions

The pandemic has been a challenging time for everyone, different states have had other guidelines and restrictions for citizens during this time, and people hold different views on pandemic management. There is complexity in analyzing longitudinal data over such a considerable period. We have proposed a way to create a pandemic profile at the state level by aggregating data for each state at the county level, from colleges and correctional facilities. We have observed statistical significance in people's sentiment towards aspects like masking and the proportion of death to cases in the state. Those three data sources are decent indicators of the spread and effectiveness of pandemic preparedness and management across micro and macro factors. We have also shown that it is possible to label states based on death the case ratio across the pandemic and detect whether a state will be in the portion of states that manage the pandemic effectively. Overall, there are distinct features that can predict how a state will manage a pandemic and how people think about specific measures that will impact the overall deathliness and spread of a pandemic.

6 Links

6.1 Submission Links

GitHub]
Databricks Notebook

6.2 Data Links

New York Times Data Set

References

[1] *Home*. May 2022. URL: https://covidprisonproject.com/.

[2] Kaifeng Yang. "What can COVID-19 tell us about evidence-based management?" In: *The American Review of Public Administration* 50.6-7 (2020), pp. 706–712.

[3] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=S1zk9iRqF7.

Appendix

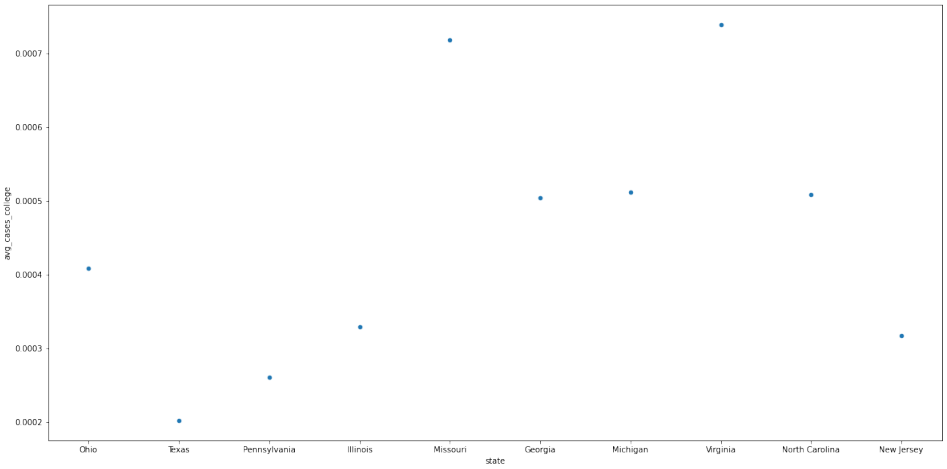


Figure 1: State x Average College Cases

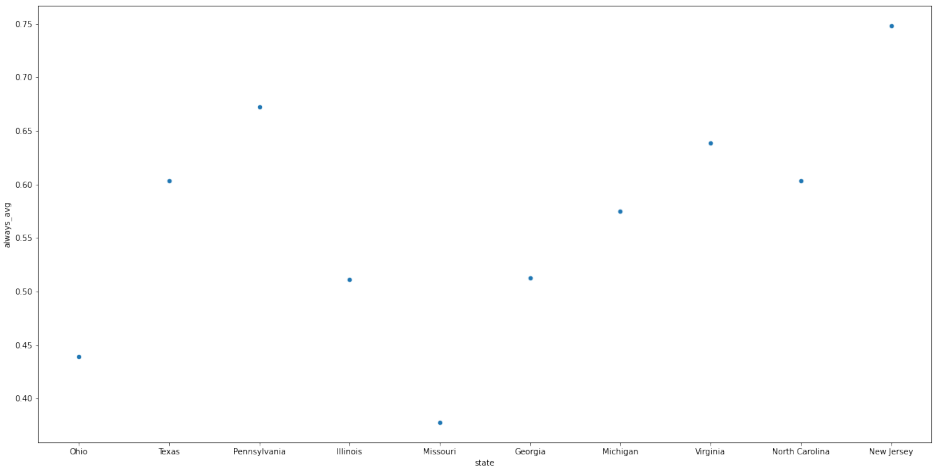


Figure 2: State x Always Mask Sentiment

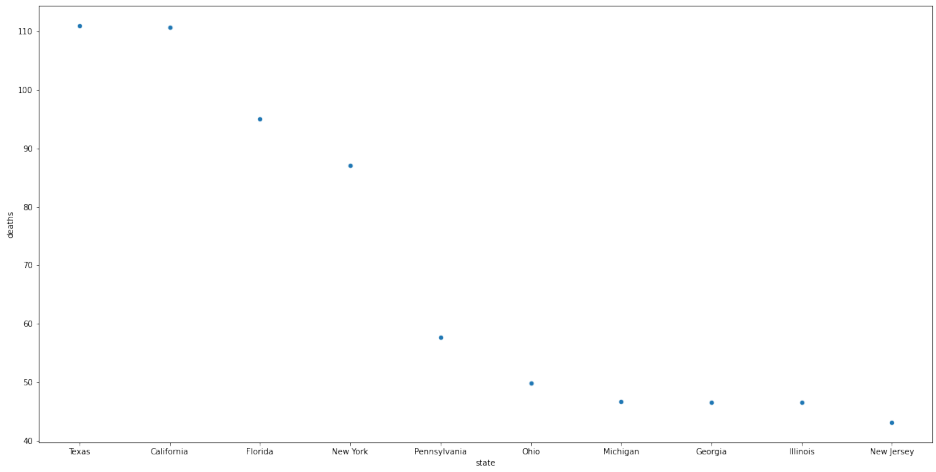


Figure 3: State x Deaths Statistic For Labeling

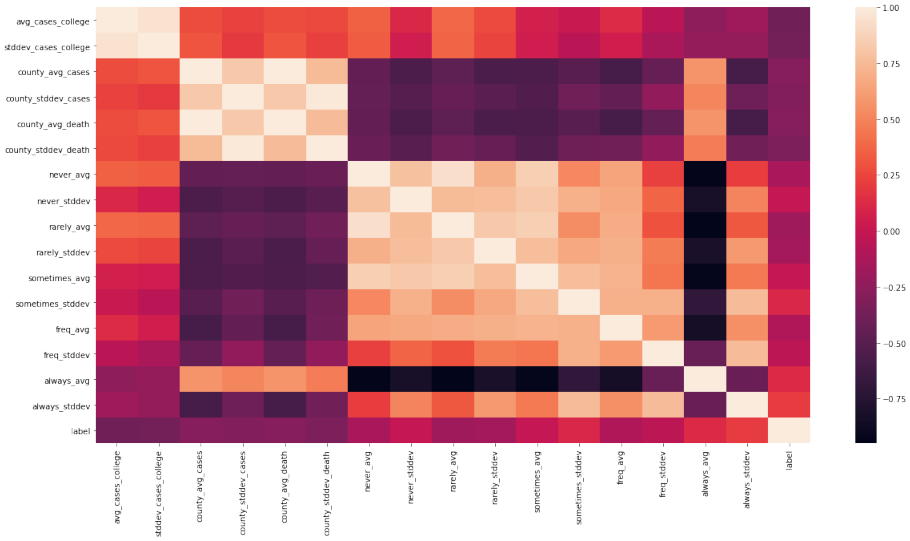


Figure 4: Pearson Correlation Among Feature Set