COVID-19 Dataset Report

by Yin Yin Teo and Christine Lu
University of California, Berkeley

Abstract

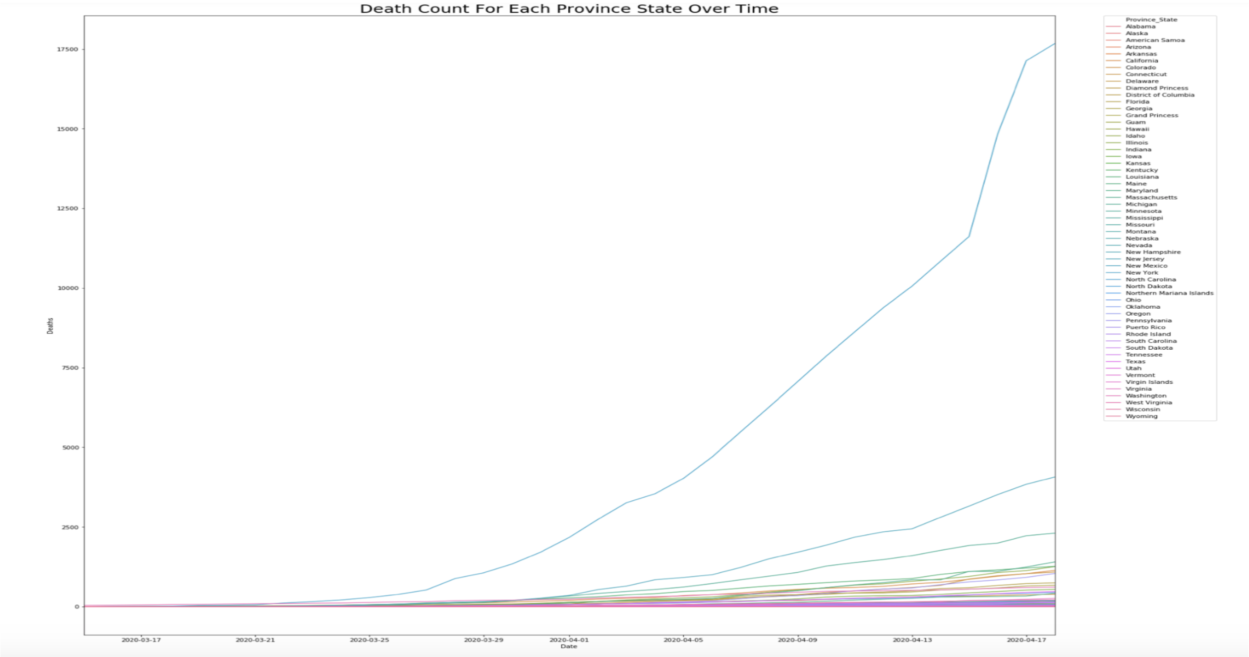
The worldwide impact of the COVID-19 pandemic, specifically in the United States, has led many data scientists and researchers to analyze patterns and provide insights on factors that may increase mortality rates and other individual health conditions that may elevate the likelihood of death for an individual. Thus, as data science students, we seek to provide some answers to these questions by analysing the COVID-19 Dataset in this project, which includes four different tables containing information about COVID-19 cases based on state, time, deaths, and cases in the United States. What kind of features would most heavily impact the survival rate of someone diagnosed with COVID-19 and how accurately can we predict death counts based on those features? By developing models with Lasso Regression and utilising methods like one-hot encoding (OHE), data cleaning, and other skills from our Data100 class, we aim to provide some clarity and precautions for those who may have underlying health/medical conditions or may be living in specific states where mortality rates are higher.

Introduction

With more than 1.4 million confirmed cases and almost 84,000 deaths, the United States has become one of the global centers for the global COVID-19 pandemic. The initial outbreaks in China and Italy have slowly given way to a nationwide public health catastrophe. 95% of Americans are under some form of stay at home order, and the gravity of the situation in the US has drawn the attention of public health experts, infectious disease experts, and policy makers from across states as well as the federal government. The COVID-19 Dataset, compiled and cleaned by the Yu group at UC Berkeley Statistics and EECS, contains information about hospital and county-level data from public sources and serves to support data science efforts to combat the COVID-19 virus. In this project, our objective is to predict death counts on 4/18/2020 by finding the ‘best’ features to model our predictions.

Description of Data

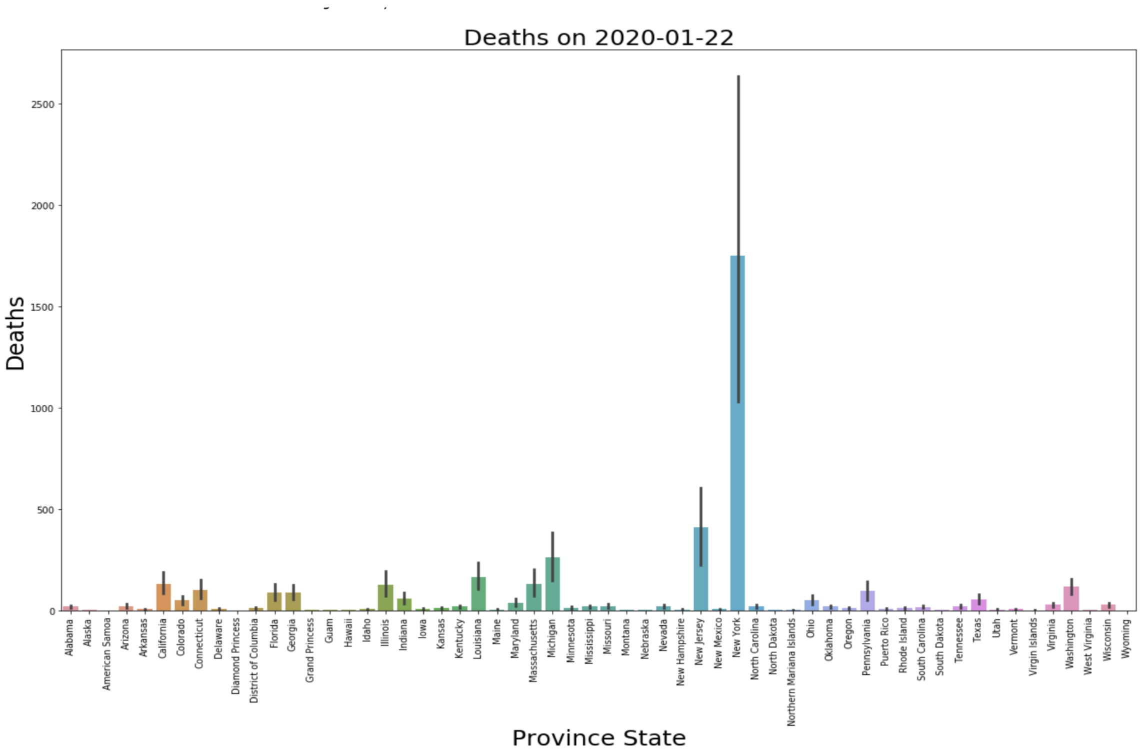
The data provided contains information on hospitals, states, counties, confirmed cases, and confirmed deaths in the US. We decided to utilize the data at a state level, which included COVID-19 cases and deaths from USA Facts and NYT as well as COVID-19 health risk factors, health resource availability, and other information about the general population such as medical history and age. First, we decided to take a look at the data through each state in the US and the number of deaths recorded for each state. In order to do this, we dropped all of the NA values in the ‘time_series_death’ table and grouped by ‘Province_State’, aggregating each column by the sum to find the total number of deaths for each state. We were interested in showing the development and increase of COVID-19 cases for each state throughout time, specifically from 3/15/2020 to 4/18/2020, so we decided to create a line plot. However, we realised that the ‘Date’ column was not standardised to the correct date format, so we converted all of the values in the column to datetime format in order to plot datetime as a continuous variable along the x-axis. We then created another table ‘graph_deaths’, which contained only the ‘Province_State’, ‘Date’, and ‘Deaths’ as columns. The graph created is as shown below:



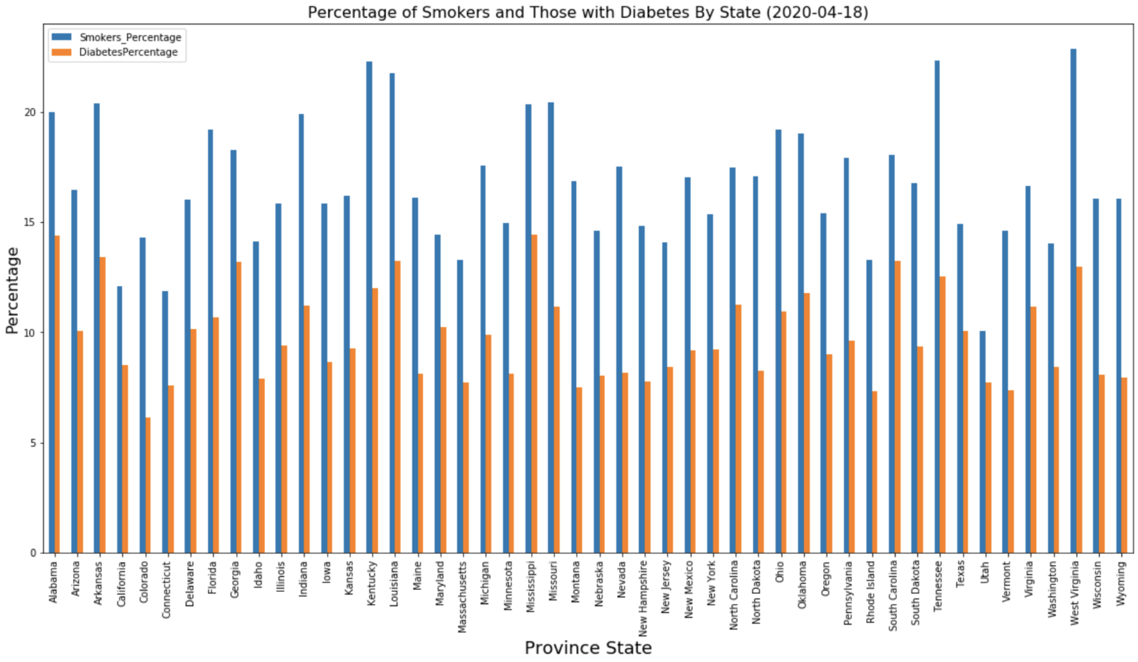
(Figure 1)

We used the *hue* feature to colour-code each state with a unique colour, and noticed that New York and New Jersey (blue) seemed to have higher rates of death compared to other states like Virginia and Wisconsin (purple), as shown by Figure 1. This made us question why New York and New Jersey specifically had a higher death rate, and by taking a look at our ‘graph_deaths’ table again, we quickly found that both states seemed to have higher populations. Nevertheless, there could still be other factors that may contribute to high mortality rates in certain states, so we decided to take a look at some of the features for each individual state that may contribute to mortality rates for one specific date: 4/18/2020.

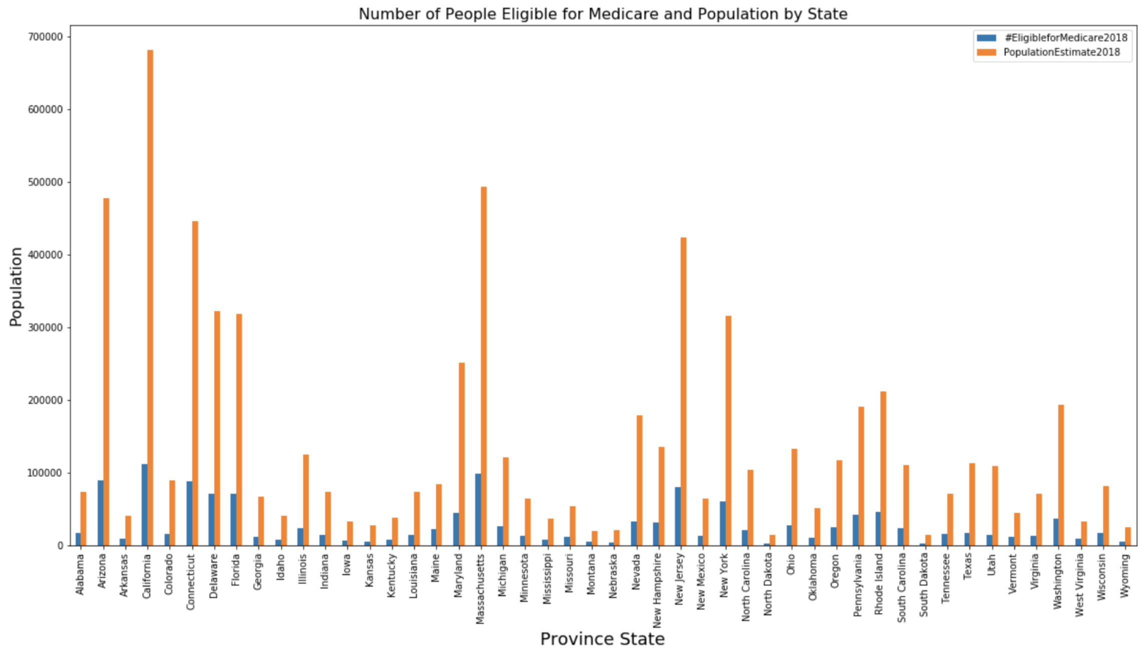
We also created other plots to further analyse the data, like *Death vs Province State* (Figure 2), *Smokers/Diabetes vs Province State* (Figure 3), *Number of People Eligible for Medicare vs Province State* (Figure 4), and *Death Rate per 100,000 People vs Province State* (Figure 5):



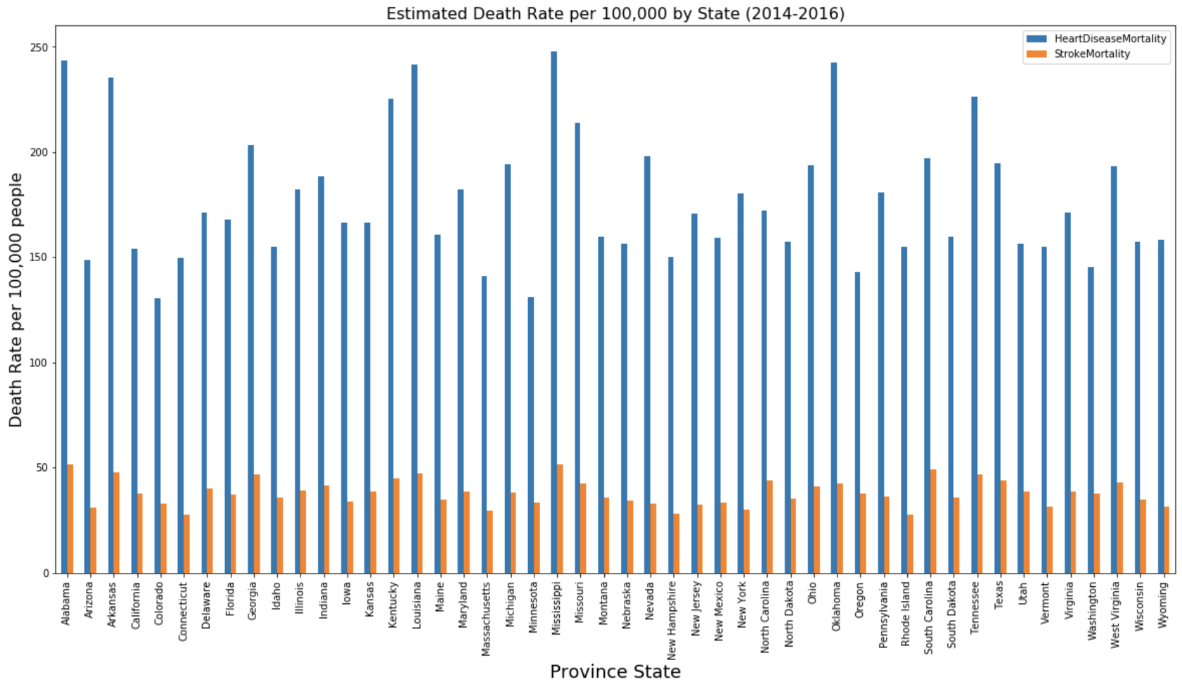
(Figure 2)



(Figure 3)



(Figure 4)



(Figure 5)

The first part of data cleaning involved taking a subset of the county data with the specific features we chose based on what we thought would affect the growth of mortality rate, named ‘county_subset’. These include ‘State’, ‘Smokers_Percentage’, ‘DiabetesPercentage’, ‘HeartDiseaseMortality’, ‘StrokeMortality’, ‘#EligibleforMedicare2018’, ‘MedianAge2010’, and ‘PopulationEstimate2018’. Then, we created a new column called ‘MedianAge2018’ which was the ‘MedianAge2010’ added by 10 to get the median age for the year 2018. 200 rows were dropped and two rows associated with the state of South Dakota were also dropped, both due to NaN values.

Potential features that contained more than 20% NaN values for the specific state were also dropped and other features that were uniform values for all data points including ‘federal guidelines’ for 3/16 and ‘foreign travel ban’ 3/11. We then merged ‘county_subset’ with the deaths table in order to combine deaths on 4/18/2020 to the features table, and named the new table ‘joined’. We also utilised one-hot encoding (OHE) for the each of the states in the ‘Province_State’ column in order to convert those categorical variables into a form that would allow us to use algorithms for predictions, where each of the states were transformed to a numerical value in the final table, ‘graph_deaths’, and used for further analysis throughout the project.

Description of Methods

Since we wanted to predict the mortality rate for the latest date available in the dataset, we thought the proper method was to select several features from the given datasets that would give the highest accuracy using techniques of feature engineering. From Piazza discussions and lectures, we knew that disease prediction models similar to SIR would be similar to the models we were looking for. We also knew Linear Regression would be an inaccurate model to predict deaths, as deaths cannot be predicted by a line, as evidenced by the curve of the SIR models. From previous labs, homeworks, and projects, we saw that other models used for feature engineering included Logistic, Lasso, and Ridge Regression. First, we selected a couple features we saw from the multiple data that we thought would affect mortality rates the most, such as the amount of people eligible for medicare, the median age of people, and population estimates for certain regions. Since Logistic Regression doesn’t include a penalization for overfitting, we decided to choose either Lasso or Ridge Regression. Ultimately, we decided to use Lasso Regression to fit a model, since it gave us a lower validation error than Ridge Regression.

For choosing which features to incorporate into our model, we first wanted to see how much including states as features as OHE features would affect the model accuracy by testing the accuracy of a model including some random features compared to the accuracy of a model with those same random features plus states one hot encoded as features. States seemed to make the validation error too low to be true, and we were afraid of it creating a bias for our model, so we decided to not include states as a potential feature for our model.

We decided to standardize all potential feature columns in order to perform lasso regression. In lecture, we learned that when applying Ridge or Lasso Regression, normalizing features is important. This is because having features that have very different ranges of values cause the magnitude of coefficients in the regression to differ greatly. By normalizing all of our features, we remove the variation in coefficient values in the regression due to difference in ranges of our features (so that each variable is regarded equally and the different coefficients in front of each feature is not due to the different magnitude of values).

We decided to choose around 6-10 features based on past labs and homeworks, so as to avoid the possibility of overfitting or having heavily dependent features, as shown in the "pitfalls of feature engineering" lecture. Assuming that each feature chosen would be independent of any other feature in the set of possible features we created, we used an algorithm that generated a random number of features randomly, and for each list of features we computed the cross validation score of a randomly selected X_train (consisting of columns corresponding to the features) and the associated y_train (consisting of the deaths corresponding to the X_train values). We then looked over the list of features that produced the lowest cross validation error across all samples. We reran the random algorithm a couple times to see which features were most often in the list of features with the lowest cross validation error, and used those features to develop our final model. We then tuned our hyperparameter alpha for Lasso Regression by testing several alphas on our final model using a loop over many alpha values across a certain range (a strategy from lab08), and found the ideal alpha. Lastly, we tested our model on the test set and found our test root-mean-squared-error.

To avoid introducing more bias or confounding factors, when generating test and train sets, we created the same consistent pseudo-random train and test rows using sklearn's train_test_split function. We also used the same alpha for Lasso Regression across all models we created (before we tuned alpha at the end). We used an initial alpha value of .01 and a test-train split of 20-80, since those are most often used in practice.

Summary of Results

We found that the population of females aged 65-74 in 2010, indicators of location, the starting date of 3/30/2020 against >50 gatherings, the starting date of 3/30/2020 for entertainment and gym closings, the population of males aged <5 and 5-9 in 2010, the 3/19/2020 starting date of the prohibition of restaurant dine-ins, and confirmed cases for certain regions as features best predicted death rates on 4/18/2020. This leads us to believe that the gender and age of the population, location, and the relative number of cases and hospitalizations, as well as government issued policies about social distancing tend to affect death rates. Although the composition of the population can't be changed and the government can't make policies to change that, creating earlier policies encouraging social distancing (the closing of public facilities for example) has proven a useful feature in differentiating regions by their number of deaths by covid-19, and may lower the death rate if initiated.

Discussion

Some important features include the population of older females and the population of younger males. It was interesting how the model weighted older females and younger males higher than other gender or age features. We also found it interesting how much using state (as one hot encoded columns) or location features affected the accuracy of the data; as a result, we decided not to use states as a feature.

We thought that features such as the percentage of smokers, diabetes, heart disease mortality, and stroke mortality would be important in predicting COVID-19 death counts since many diseases have in-common risk factors, yet our model rarely chose those features compared to features like age or government policies when we ran our algorithm multiple times.

Some challenges we faced included framing our initial problem. We realized that it would be difficult to create a model predicting either death rates or confirmed cases over time with static features given to us in the dataset (and we thought that using some sort of regression on previous dates to predict future date rates or cases would be too simple). We ended up choosing one specific date to predict with our static features. We were also confused about how to interpret some of the format of some of the columns of the data, such as those in the gregorian time format. Using help from those on Piazza who had similar issues, we were able to extract timestamps from some of the columns, and were able to more easily interpret the values of those columns.

For the analysis of features, we chose features rather randomly; therefore, we made the assumption that any random combination of features could equally properly predict the mortality rate. We assumed that the features are independent of each other, although many features in the given data sets seem to be correlated. For example, some features are just the same feature measured over time. In lectures and in homeworks, we were shown how having dependent columns and redundant features could make the matrix not full rank, and create an unsolvable model, so this could be a potential problem with our analysis. Another problem with random generation and similar features could be that changing any hyperparameters or parameters (such as alpha in Lasso Regression, or the size of the test-train split, random states) changes which features appear most often when reducing cross validation error, because it seems that many of the potential features were effective at predicting death counts. In addition, the randomness of the algorithm itself makes it so that over every run of the algorithm, it’s almost certain that you won’t get the same set of features that minimizes cross validation error.

Some of the ethical concerns we noticed with the COVID-19 dataset are issues with representation and the question of whether or not informed consent was given for inclusion in the dataset. The representation may be skewed as the dataset is not the raw data, which means it did not exist prior to argument or interpretation. Instead, it had already been cleaned, compiled, and documented by the Yu group at UC Berkeley Statistics and the EECS department. This means that choices had already been made about what kind of data to collect and how the data should be presented. For example, demographics, health resource availability, health outcomes and risk factors, social distancing and mobility (private data), and voting ratio (Democrats:Republicans) are all different methods on how the COVID-19 data is being represented. However, there may be other factors that were not included and therefore not represented in the dataset, which is an ethical issue as bias may be introduced. Some of the data collected also seems to be personal or private data, which draws upon the question of whether or not informed consent was given before letting the data out for public use. Even though this information was not utilised in this specific project, the dataset included information on social distancing and mobility data, which was marked as private data. Other categories such as personal health history and medical conditions may also be considered private data. Voting ratio also discloses personal information such as political affiliation and voting choices, which is also considered private data. Hence, making sure the data was acquired correctly without violating the privacy of individuals becomes an important part of making sure the data is being used ethically.

Some additional data that would strengthen our analysis would be the location of each confirmed death case, which would allow us to track and cluster such cases and allow us to track linked COVID-19 cases to a specific individual. Other health conditions such as those with lung disease, asthma, cancer, bone marrow or organ transplantation, immune deficiencies, AIDS or HIV previously, severe obesity, or other chronic liver and kidney diseases may also help determine the mortality rate of an individual who has contracted COVID-19. It would also allow us to form other hypotheses based on specific health conditions such as these and if they would contribute to a higher mortality rate as compared to the current features available to us now.

Other ethical issues with studying the specific problem of what features may contribute to higher mortality rates in different states is the idea of classification. Some of the data pertaining to demographics include gender, and we noticed that there were only two categories: male and female. However, since those are not the only two categories that exist in today’s society, there may be an ethical issue here since identity is an important relationship between data and the individual. Instead of disregarding other gender categories, the dataset should either be adjusted to address those concerns or there should be a note at the end of any analysis recognising the disincluded categories. Another ethical concern is agency, which is the ability to say no to having one’s data collected. Some of the categories like party affiliation and medical records as mentioned earlier may or may not have allowed the option of agency. Thus, these concerns should be addressed by making sure the data was collected ethically or considering disregarding these types of data to avoid abusing unauthorised and unethical usage of personal data.

We think the randomness of our approach is good so as to not introduce any bias, and the frequent random generation of features as well as the random generation of the number of features to include in our model also helped us find the ideal categories of features and number of features to fit our model to not overfit/produce variance. Using Lasso Regression also seems to create a pretty accurate model.

In all, we were surprised that the features we had chosen based on intuition were somewhat accurate, even though they were not as accurate as the randomized sampling. In the future, we could attempt a SIR model and incorporate the progression of dates and times to predict how many cases of COVID-19 there will be in the future or even the mortality rates across specific states. We could also incorporate other COVID-19 datasets not included this time to provide more accurate predictions and conclusions and maybe even make our findings available to the public to keep them informed and aware for further transparency on the issue.