University of California, Berkeley

MODELING COVID-19:

PREDICTING SOCIAL VULNERABILITY INDEX THROUGH INFECTION RATE ACROSS US COUNTIES

Abdul Choudhry, Phuong Thao Nguyen, Sanjana Shah

DATA 100

Joey Gonzalez, Ani Adhikari

13 May 2020

Abstract

In this paper, we develop several models to predict how social distancing affects the spread of COVID-19 mainly through measuring the social vulnerability index across United States counties. We analyze how historical healthcare trends and social distancing efforts enacted by individual counties and states nationwide play a role in the virus' spread by only using features that counties nationwide have released in the past including but not limited to social vulnerability index percentile, number of ICU beds available for each hospital, as well as mortality demographics. Through our analysis and results, we hope to answer the following two guiding questions: How have social distancing efforts affected the rate at which the virus has increased or decreased its spread? What does the social vulnerability index tell us about how prepared each respective county was in response to COVID-19 preparedness? We analyze potential answers to these broad questions and propose results through classification techniques, cross-validation, and models including Logistic Regression, Linear Regression, Decision Trees, and Random Forests.

Introduction: Data Cleaning

In order to perform our analyses to the aforementioned questions, we initially had to clean the datasets provided so that we were able to work with them and create some initial visualizations to further dig into our questions. The four CSV files that we were given included a combination of categorical and numerical variables and some of these numerical variables had missing values that we decided to either clean or drop entirely from the data frame. We loaded each file to get a better sense of how we wanted to use the data and after cleaning the data we thought was useful within the scope of our research questions, we created visualizations.

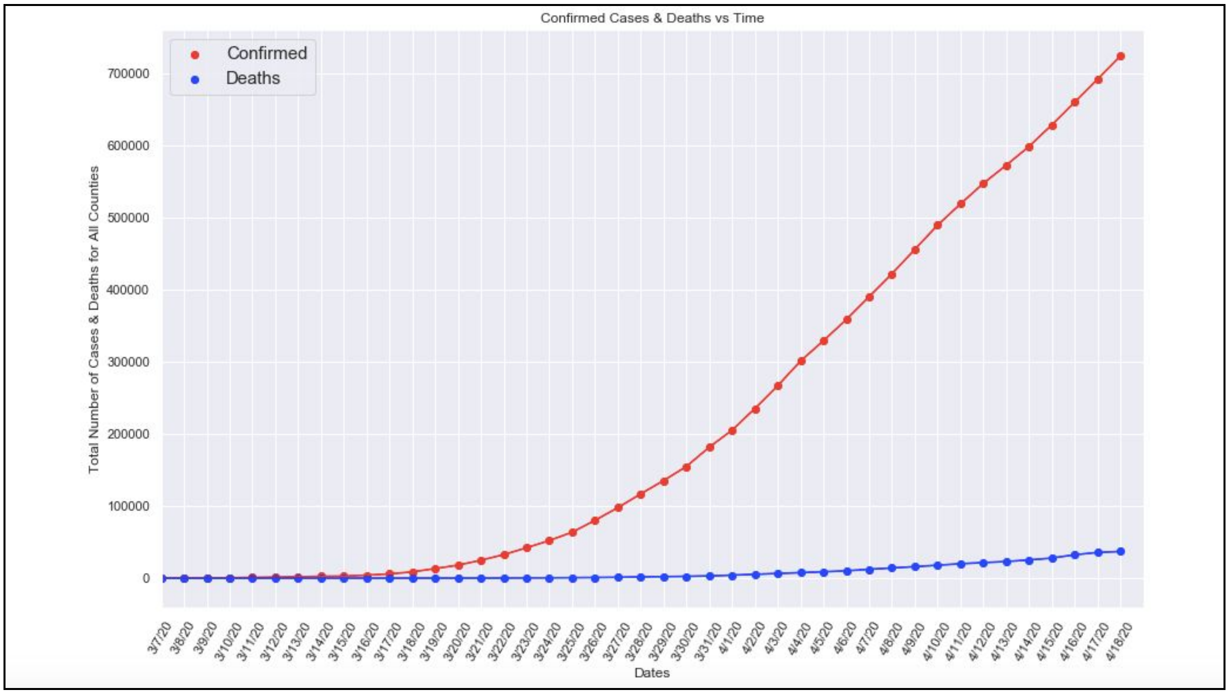
Immediately, we noticed that the 'abridged_counties.csv' file contained a lot of information regarding geography, demographics, and health resource availability on the county-level for the US. In order to work with this dataset, we first dropped all the rows with NaN values in the 'countyFIPS', 'STATEFP', and 'State' columns. This was done because the county Federal Information Processing Standards (FIPS) code is a unique numeric code for each county and we aimed to drop all rows that were not US counties. We then converted their numerical values from floats to integers to keep track of each US county thereby reducing future errors. We removed 0's from the beginning of the counties' FIPS code, as we noticed that other data frames listed county FIPS codes without the 0, and thus merging in the future would become easy. In order to better analyze the spread of COVID-19 and counties' responses, we used the CDC's 2018 public SVI Index CSV file which provides more information about the relative vulnerability of every U.S census tract. For context, the SVI ranks these tracts on 15 social factors including disability, unemployment, minority status, and many other factors that contribute to each county's SVI, which falls within the range of 0 to 1, 0 being the lowest vulnerability and 1 being the highest. Then, we look at the uniqueness of each column to scale the numbers. The number "-999" and "-999.0" were used as replacements for missing information. We replaced all values of "-999", "-999.0", and NaN values across all column values with the arithmetic mean in each respective column, and converted all floats to integers. We did not want to replace empty values with zeros because that would influence our data.

In order to address the question regarding what the social vulnerability index tells us about each county's preparedness to COVID-19, we imported 'SVI2018_US_COUNTY.csv' which contains socioeconomic, household composition/disability, minority status/language, and housing/transportation information per county. We merged this data frame with 'svi_inner' to create 'svi_merge', a data frame that not only details the social vulnerability of each county, but also has COVID-19 also has confirmed cases data for each county. We performed the same operation of dropping NaN values on this data frame to avoid missing/empty data going forward. We will use this data frame to extract the top features that impact each county's vulnerability the most and thereby predict each county's SVI Percentile to address preparedness.

We repeated this process of dropping rows with no information across the '4.18states.csv' file, which provides COVID-19 data as of April 18th, as well as the time-series data in both the time series confirmed and deaths CSV files. We also filtered out states and counties that were "unassigned" and "out of state" in these time series files in order to accurately visualize what our data was showing us. Since Hawaii and Alaska do not have counties in the 'abridged_counties.csv' file, we dropped rows for them as well. After cleaning our data and making sure no outliers existed within all the CSV files, we performed an inner merge on the time series confirmed and county data frames to keep track of the confirmed cases for each county. Subsequently, we created functions that converted the proleptic Gregorian ordinal dates within certain columns of this newly merged data frame, such as the 'stay at home' column which showed the date where each county took measures to mitigate the spread, and therefore named this data frame 'sth_inner.' Finally, before moving on to the feature selection and modeling process, we standardized our selected features to ensure that all numerical variables are within the same scale to ensure the highest possible training accuracy for our model.

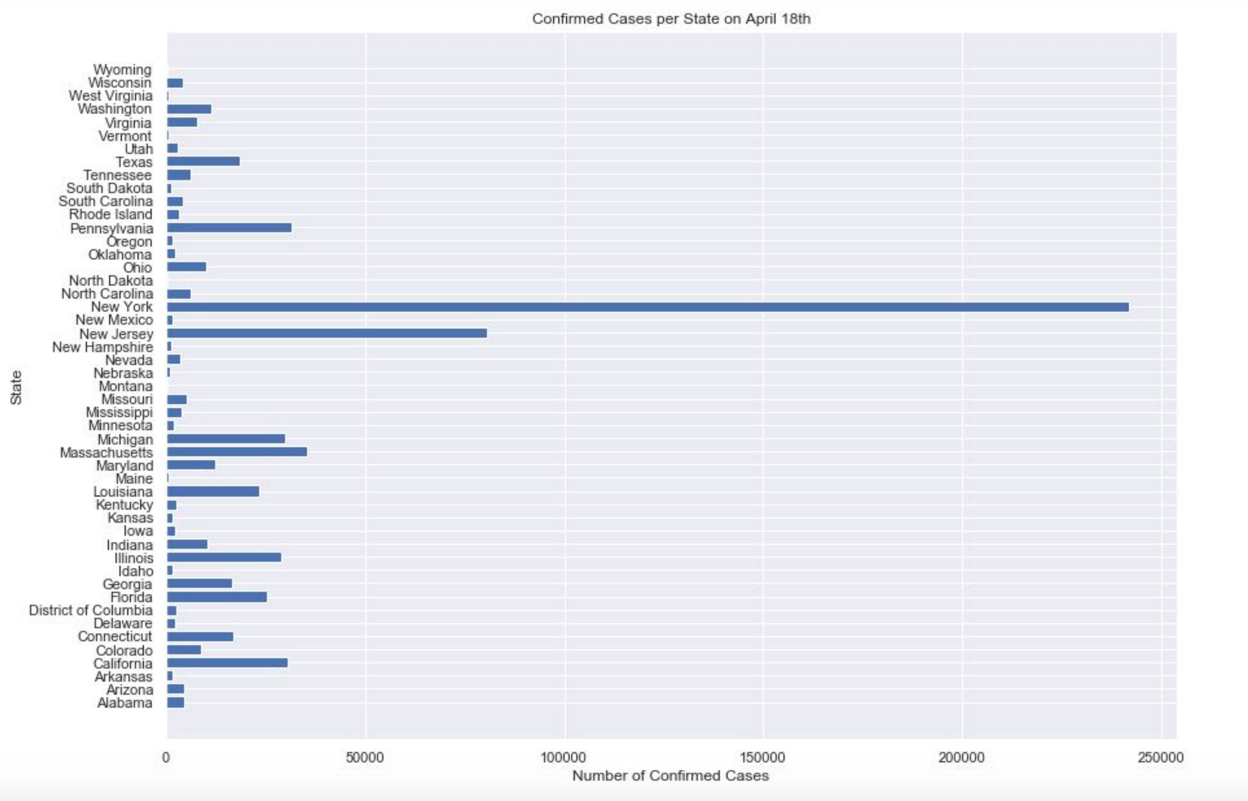
EDA & Data Visualizations

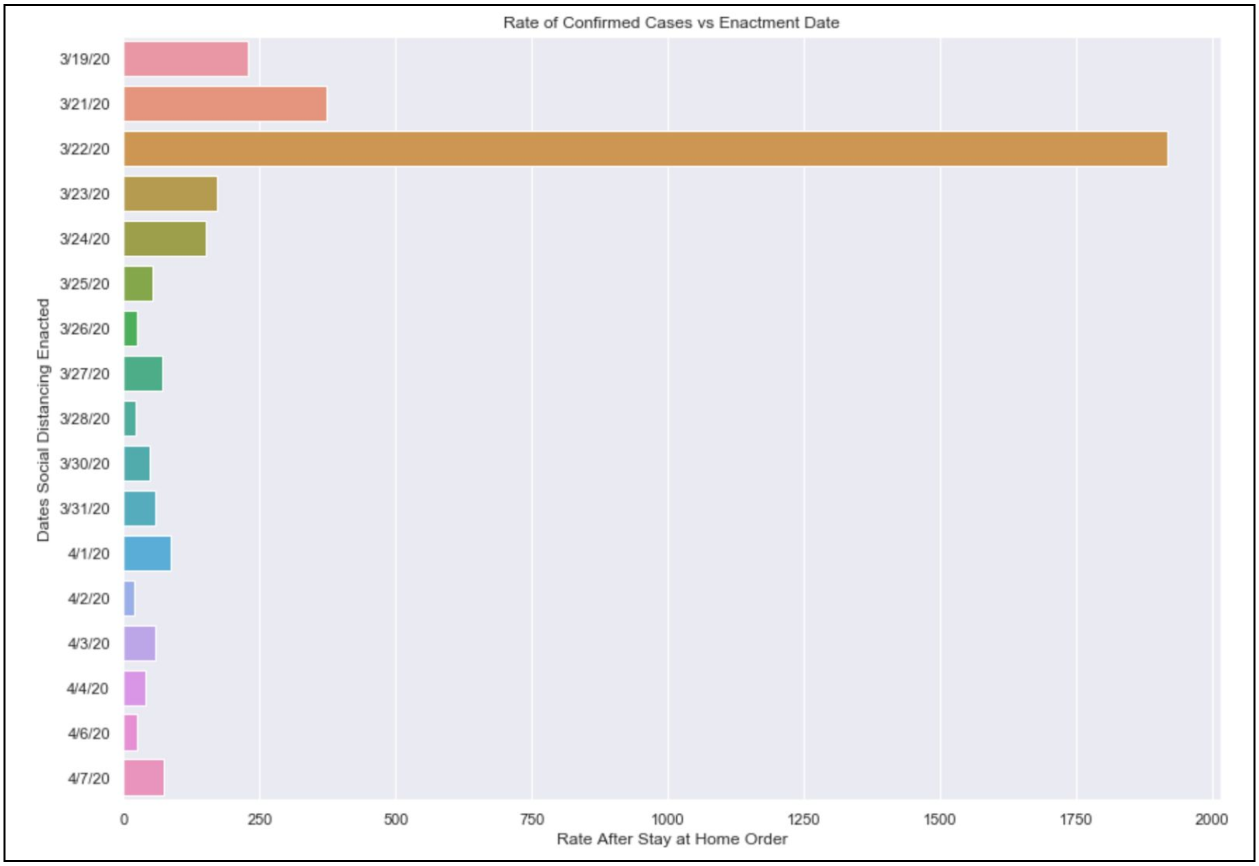
In order to initially explore the data given to us, and further explore the question of how social distancing efforts affected the rate of COVID-19's spread, we analyzed the relationship between the number of confirmed cases and the number of deaths with respect to a time interval of interest to us, which was from March 7th to April 18th. We focused on this time interval because we noticed that most cases and deaths in the US started to occur in the months of March and April, which accurately falls along the timeline of when our semester moved to virtual learning and the situation worsened. We used both time series data frames and MatPlotLib to create the following visualization:



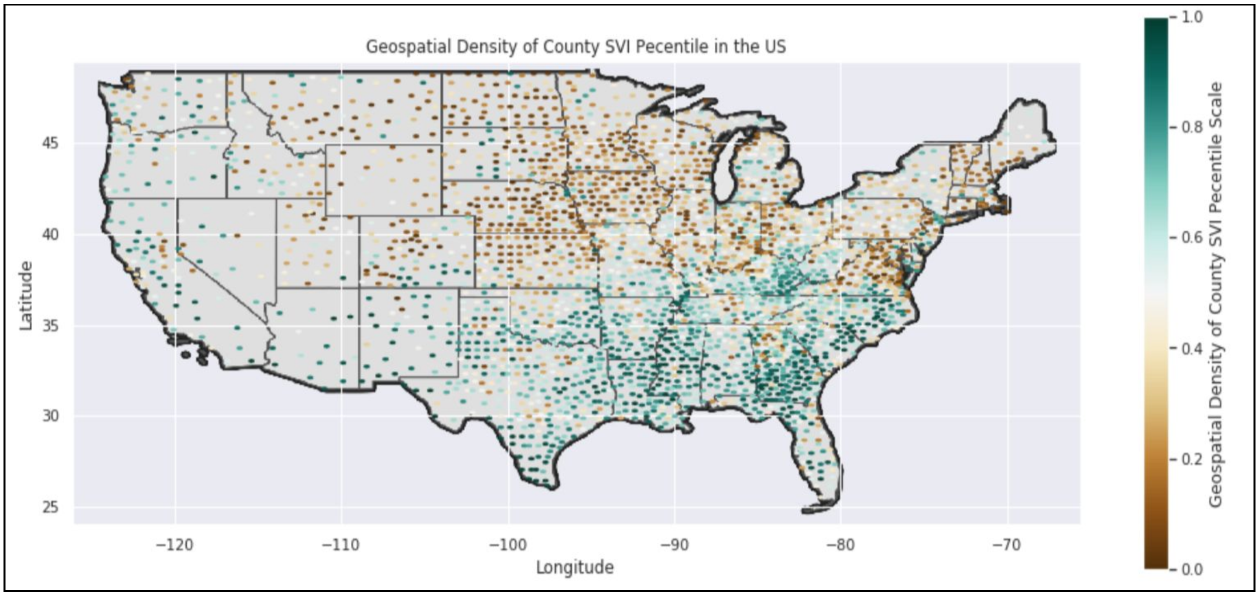
According to Ben Westcott's CNN article titled "April 18 coronavirus news", on April 18th, there were 732,197 confirmed cases in the US and 38,664 deaths (Westcott) which matches our above graph and exploratory analysis of the data given. After creating this visualization, we wanted to observe how many confirmed cases there were in each state and not surprisingly our horizontal bar plot showed that New York had the highest number of confirmed cases, almost nearly 250,000, as of April 18th.

The next step towards analyzing the spread of COVID-19 was seeing what effect the dates at which social distancing orders were enacted by each county had in terms of the average rate at which the confirmed cases increased per county. In this step, we used the function we created to convert Gregorian time to the standard date in mm/dd/yy format and created a list of the average confirmed case rates before and after a specified date. Our visualization shows that across all counties in the US, the rate at which confirmed COVID-19 cases spread decreased significantly after March 22nd, which is around the same time period when counties across the Bay Area received shelter-in-place orders. Omar Perez's online article confirms that California saw a 40% decrease in average distance traveled and the graph in this article shows that this occurred around March 23rd (Perez), which supplements the following visualization we produced:

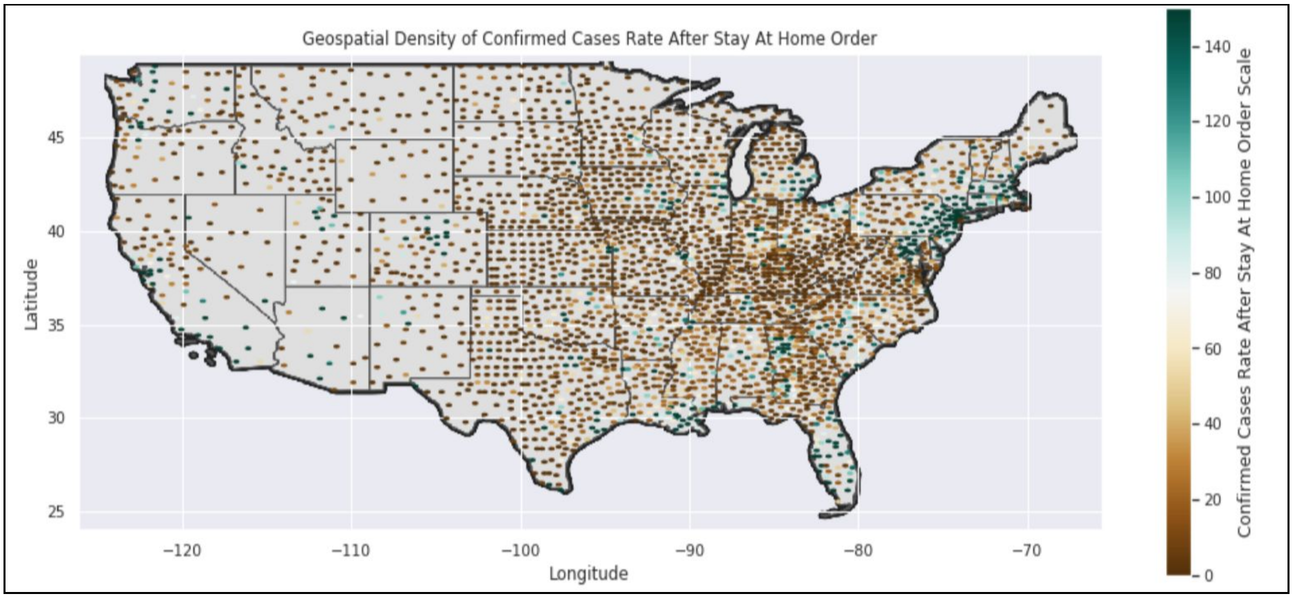

Confirmed Cases per State on April 18th
State
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Georgia
Florida
District of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alabama
0
50000
100000
150000
200000
250000
Number of Confirmed Cases


Rate of Confirmed Cases vs Enactment Date
Dates Social Distancing Enacted
3/19/20
3/21/20
3/22/20
3/23/20
3/24/20
3/25/20
3/26/20
3/27/20
3/28/20
3/30/20
3/31/20
4/1/20
4/2/20
4/3/20
4/4/20
4/6/20
4/7/20
0
250
500
750
1000
1250
1500
1750
2000
Rate After Stay at Home Order

For our final step in the visualization stage of our exploratory data analysis process, we decided to create two hexbin plots; one that detailed the geospatial density of the given SVI Percentile of each county and the other to show the rate of confirmed cases after a stay at home orders were enacted. The former plot is used to have a visualization to compare our predictions against. Using what we see in terms of each county's vulnerability, we can compare against the latter plot of each county's rate of confirmed cases to see if we notice anything surprising.


Geospatial Density of County SVI Pecentile in the US
Latitude
45
40
35
30
25
Longitude
−120
−110
−100
−90
−80
−70
Geospatial Density of County SVI Pecentile Scale
1.0
0.8
0.6
0.4
0.2
0.0

In this plot, each point on the US map represents a county and the color corresponds to its respective SVI Percentile. The browner the point, the lower it is on the Geospatial Density Scale, and therefore the lower the Social Vulnerability Index. This means that the county is less socially vulnerable and less likely to be affected by COVID-19. Whereas, the bluer the point, the higher it is on the Geospatial Density Scale, and therefore the higher the Social Vulnerability Index. This means that the county is more socially vulnerable and more likely to be affected by COVID-19. Following this, we can see that most counties located in the Northeast of the US tend to have a lower social vulnerability, and therefore we can predict that they will be least affected due to COVID-19. On the other hand, most counties located in the Southeast of the US tend to have the higher social vulnerability, and therefore we can predict that they will be most affected due to COVID-19. To explore this, we created a hexbin plot of the average rate of confirmed cases per county after their respective stay at home order:

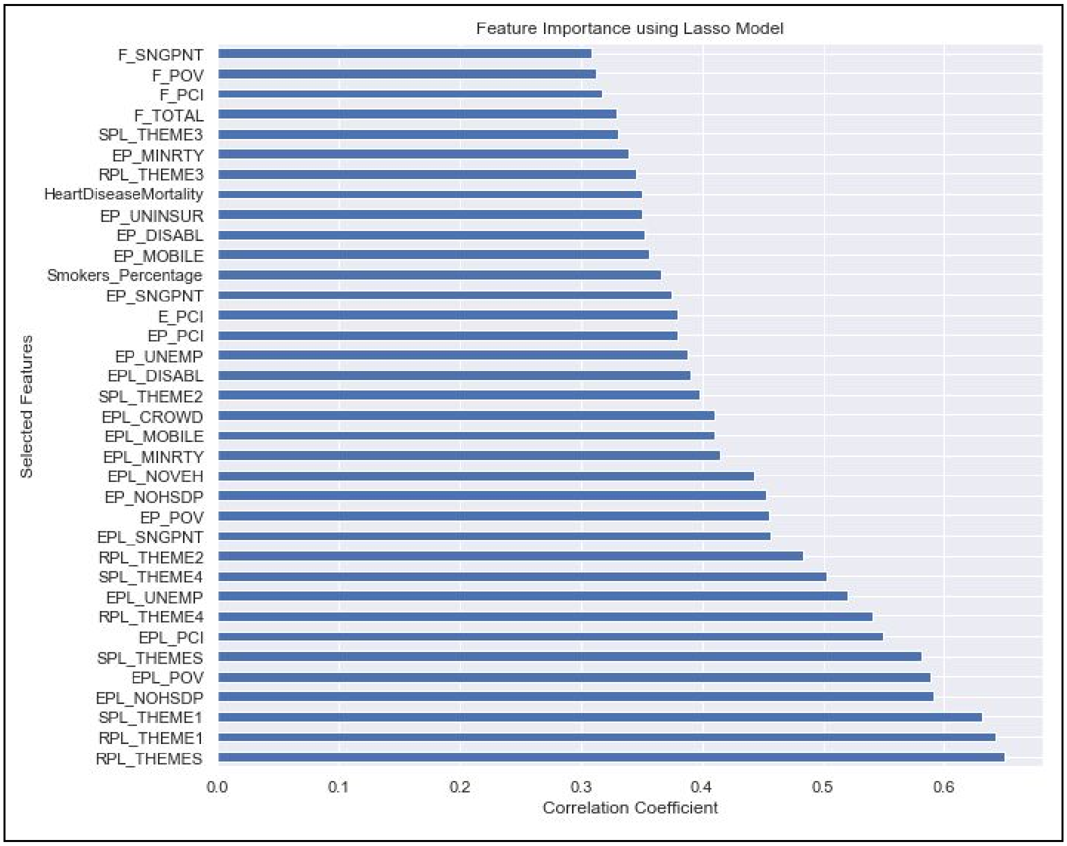

Geospatial Density of Confirmed Cases Rate After Stay At Home Order
Latitude
45
40
35
30
25
Longitude
−120
−110
−100
−90
−80
−70
Confirmed Cases Rate After Stay At Home Order Scale
140
120
100
80
60
40
20
0

In this plot, each point on the US map represents a county and the color corresponds to its respective rate of confirmed cases after date of stay at home order enactment. The browner the point, the lower the rate of confirmed cases for each county, and therefore the less it was impacted by COVID-19. The bluer the point, the higher the rate of confirmed cases for each county, and therefore the more it was impacted by COVID-19. Surprisingly, most counties in the Southwest, though were shown to be highly socially vulnerable in the previous figure, were not largely affected by COVID-19. However, those in the Northwest that were not highly socially vulnerable, seem to have a higher rate of confirmed cases of COVID-19. This leads us to the following questions: is there any correlation between the SVI Percentile and the rate of confirmed cases, and what factors determining SVI Percentile would be best to predict the correlation between SVIPercentile and the rate of confirmed cases in each county?

Method, Modeling, and Experiments

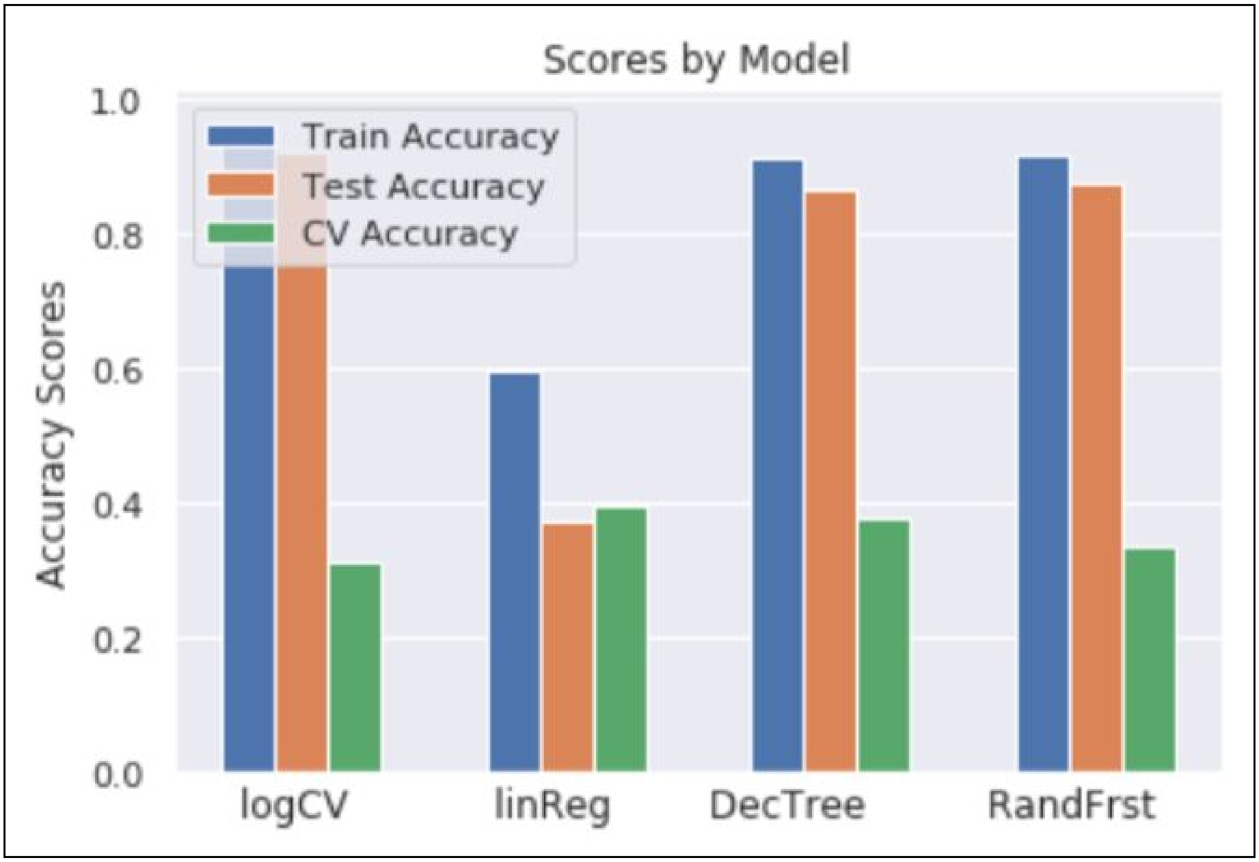
In order to implement our feature design and modeling process, we initially selected all the numerical features from our merged ‘svi_merge’ data frame, which contained both COVID-19 related healthcare conditions and the CDC’s county data including social vulnerability features such as socioeconomic, household composition/disability, minority status/language, and housing/transportation. Once we selected these combinations of features, we dropped the features that included proleptic ordinal dates to solely focus on features that we thought would better predict the SVI percentile of each US county, emphasizing the use of COVID-19 features within this data frame. Since SVI percentile is our response variable, we created a binary series of 1s and 0s, where all SVI percentiles in the current feature matrix greater than 0.5 were assigned to 1 and all SVI percentiles less than 0.5 were assigned to 0 in order to classify our response variable and use this series to smoothly train our model. Using a combination of COVID-19 and CDC data, this would essentially be used to train our models and predict which counties and states in the US have the highest or lowest social vulnerability and how prepared these states were in response to COVID-19. After creating a binary series, we then standardized our current feature matrix in order to have all variables on the same scale and dropped the column we are trying to predict, which in our case is ‘SVIPercentile’.

Next, we used scikitlearn’s built-in LassoCV linear model with iterative fitting along a regularization path to execute the feature selection process for us by passing in our newly created standardized feature matrix as an argument. This regularization method falls under an umbrella of embedded methods that penalize a feature given a coefficient threshold, which our Lasso model does. Using the ‘SVIPercentile’ as our response variable, we found that the best regularization parameter alpha, which is the amount of penalization chosen by cross-validation, is 0.001175. Out of all the 278 features within our feature matrix, our Lasso model picked 46 best features and eliminated the other 228 through this iterative process. The correlation coefficient threshold that we selected was 0.3, so out of the best 46 features that were most correlated with SVIPercentile that Lasso computed from our feature matrix, we chose the top 36 most correlated features with SVIPercentile, which is shown in our visualization below:



From these features, which seemed to primarily be from the CDC’s dataset, we selected the first 20 most correlated ones and added them with time-series data of confirmed COVID-19 cases from April 10th to April 18th in order to incorporate features from COVID-19 data that we thought would be most impactful when training our models. After feature selection, we split our data into training and testing sets, 70%, and 30% respectively, using the feature matrix that had already been standardized previously. Our independent feature training matrix (X_train) included all these features using the training data frame and our response variable was SVIPercentile.

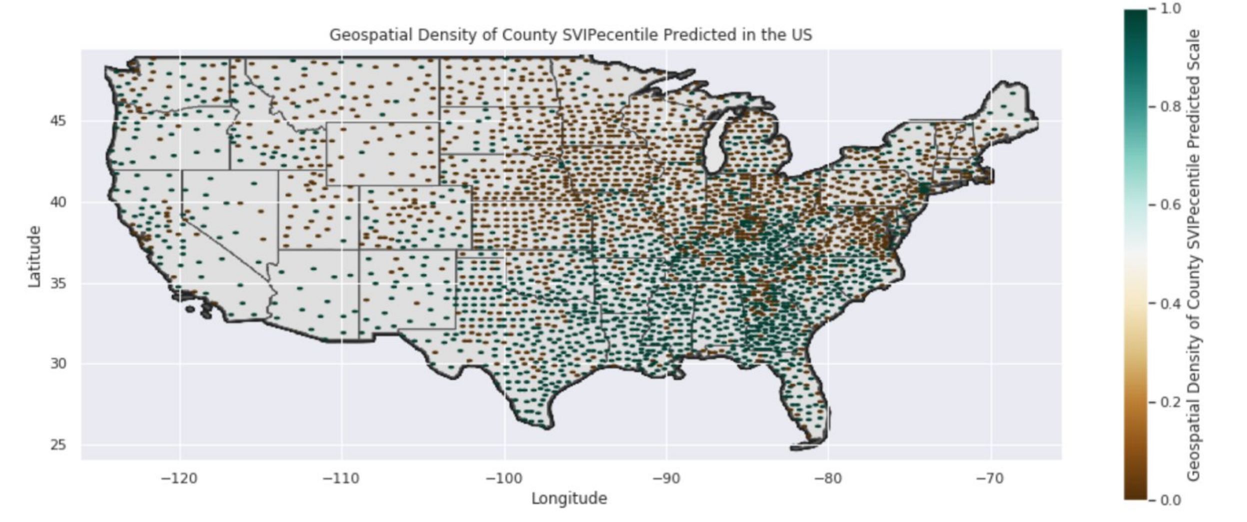
We created the following four models in order to predict the social vulnerability percentile that counties nationwide have, given COVID-19 features and CDC SVI data: LogisticRegressionCV, Linear Regression, Decision Tree, and Random Forest. In the logistic regression model, we included the "Cs" parameter to account for regularization and overfitting and used five-fold cross validation to increase our training accuracy, which approximated to about 96% using this model. When using Linear Regression, we passed in the same independent feature training matrix and response variable as before; however, did not account for cross validation and regularization which lowered this model's training accuracy, to approximately 59%. For both the decision tree and random forest models, we set the minimum number of samples required to split an internal node to 6 in order to maximize our training accuracy, which for both models was roughly around 91%. The following graph we produced shows the relationship between training accuracy, testing accuracy, and CV score across each model:

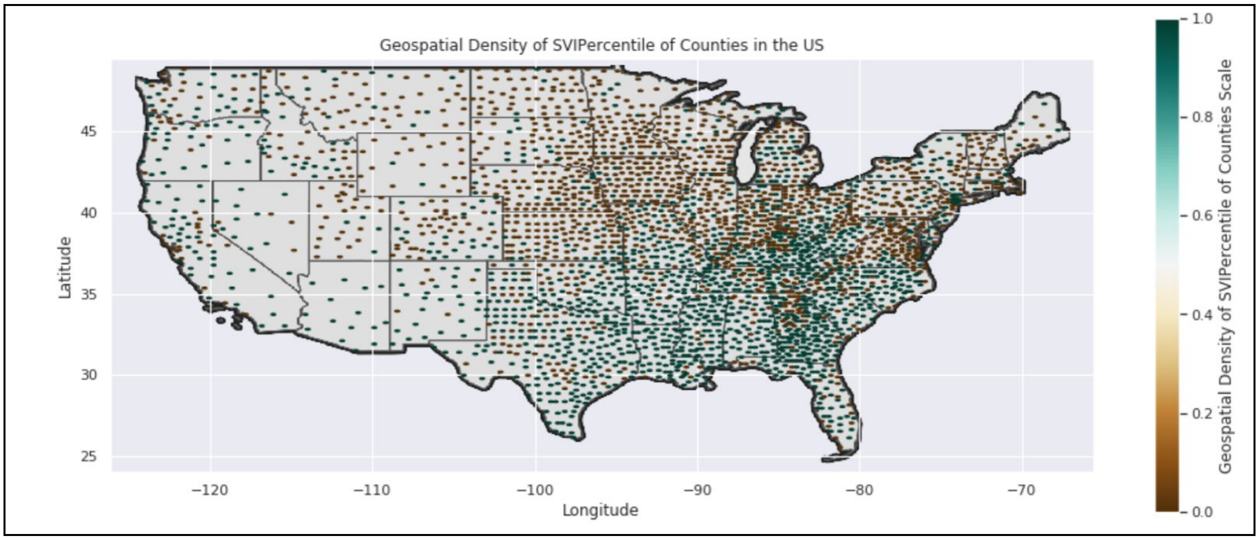


Analysis & Conclusions

What we found interesting is that there appears to be little to no correlation between the SVIPercentile, its factors, and the rate of confirmed cases. It was disappointing to see this because we would assume features like ‘F_AGE65’ (percent of people over 65) and ‘EPL_MOBILE’ (percent of the mobile home) would influence the rate of increase. In the beginning, we generated a bar graph to see the rate of confirmed cases in each state by taking the rate of each county and taking the mean. We analyzed that states with earlier social distancing enactment dates have higher rates of confirmed cases; however, they also have higher confirmed rates before the enactment of social distancing. Moreover, we cannot assume that social distancing increases the rate of confirmed cases, since this is to show that states with high confirmed cases have taken earlier measurements; with the high initial cases, especially in New York, the increase of confirmed cases overtime is drastic. Misinformation like this would be used to mislead the severity of COVID-19 to the public, which is an ethical issue. This can cause certain states and counties to have conflicts within one another in the context of economic prosperity, also creating political and ethical tensions. As of May 13th, we have not reached a peak for the confirmed cases all over the US, and therefore it is ambiguous to tell what the best precautionary measure is.

Therefore, social distancing cannot alone be the predictor of the rate of confirmed cases, so we look at variables going into the SVIPercentile of each county, with more than 30 Percentile correlation. Some of the best predictors for SVIPercentile based on our classifiers are ‘Smoker Percentage’, ‘EPL_Poverty’, and ‘EP_Mobile’. However, there is no correlation despite the hex map analysis of the SVIPercentile and rate of confirmed cases being inversely related. Even though our data shows high accuracy in the testing and training data, CV scores are low, CV scores tell the the ratio between the standard deviation and the mean. However we noticed that there are more than 2 percent differences between our test and train data, with low RSME. Since we predicted on a binary scale, our model is limited to what we can classify. In order to keep the classifier faster, we adjusted our scales, this has an influence on the specificity and sensitivity of our data. To fix that, it can be useful to use a more advanced back propagation classifier.


Geospatial Density of County SVIPecentile Predicted in the US
Latitude
45
40
35
30
25
−120
−110
−100
−90
−80
−70
Longitude
1.0
0.8
0.6
0.4
0.2
0.0
Geospatial Density of County SVIPecentile Predicted Scale


Geospatial Density of SVIPercentile of Counties in the US
Latitude
45
40
35
30
25
Longitude
−120
−110
−100
−90
−80
−70
Geospatial Density of SVIPercentile of Counties Scale
1.0
0.8
0.6
0.4
0.2
0.0

By generating hexbin plots based on our classifiers and the actual SVIPercentile, we see that our classifiers did predict accurately the SVIPercentile per county based on our features. There is definitely correlation between SVIPercentiles, socioeconomic factors, household type, minority, and household composition/disabilities, however there appears to be no correlation between that SVIPercentile and rate of confirmed cases. Therefore, we cannot assume that by being a subset of individuals in the high SVIPercentile, there will be a higher chance of getting the virus. Even though the SVIPercentile helps determine how well prepared counties nationwide are, it needs to be readjusted for a case like the COVID-19. Further investigations need to address data within the west coast of the US. Our data is limited to the middle east and the east coasts, which can only yield to conclusions drawn from these regions. Further studies can also expand on how political affiliation and preference influence the rate of confirmed cases and SVIPercentile of each county nationwide.

Works Cited

"CDC's Social Vulnerability Index (SVI)." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 2020, svi.cdc.gov/data-and-tools-download.html.

- List of columns: https://svi.cdc.gov/Documents/Data/2016_SVI_Data/SVI2016Documentation.pdf

Pérez, Omar. "Cellular Data Shows Which Bay Area Counties Are Following Stay-at-Home Orders." *KRON4*, 31 Mar. 2020, www.kron4.com/news/california-gets-a-for-practicing-social-distancing-study-finds/.

Westcott, Ben. "April 18 Coronavirus News." *CNN*, Cable News Network, 19 Apr. 2020, www.cnn.com/world/live-news/coronavirus-pandemic-04-18-20-intl/index.html.