Identifying Key Factors that Model the Spread and Growth of COVID-19 in Communities Using Computational Methods, Machine Learning, Statistical Inference, and Predictive Modeling

C. W. Chiong, R. R. Chuppala, and S. B. Banavalikar
Division of Computing, Data Science, and Society, University of California Berkeley, Berkeley CA

***Abstract*- The ongoing coronavirus pandemic has been a fairly erratic pandemic, with epidemiologists constantly updating models and predictions as medical researchers continue to gain new information about the virus. In this work, we present a feature-based, linear regression model that allows us to identify intrinsic and socioeconomic factors of a community that are most helpful when modeling the spread and growth of COVID-19 cases and deaths in American communities. We utilize computational methods, statistical inference, machine learning, and predictive modeling in order to create this model. We subsequently apply this predictive model to estimate the number of cases in 10 major United States counties, on a daily basis and analyze results. We also present a proof-of-concept exponential model that incorporates historic trends to predict the number of future COVID-19 cases and deaths and to assign a level of danger to each predicted community. Finally, we raise some ethical questions that the data brought to our attention, specifically in regards to how rural areas and blue collar workers have been left to fare with this virus.**

I. INTRODUCTION

The coronavirus pandemic is the largest global health crisis of the 21st century. However, unlike great pandemics in the past, massive amounts of data and new analytics techniques are at our disposal to help understand and combat the effects of the crisis. Allocating scarce health supplies and resources is a daunting task, and if we are able to predict where future peaks in coronavirus cases will appear, we can help prevent our health system from being overwhelmed. The purpose of this paper is to provide a quantitative analysis of how different factors ranging from socioeconomic characteristics to the average miles a person travels in a day affects COVID-19's spread and mortality to pinpoint regions of the United States most vulnerable to the COVID-19 outbreak.

Thus, we have formulated a main question and a proof-of-concept followup. What are the intrinsic characteristics of a community that are most helpful when creating a model to predict the number of future COVID-19 cases and deaths? Can we incorporate historic trends, along with our identified features, to predict the number of COVID-19 cases on a day-by-day basis? To find the best features for our analysis, we used different methods ranging from Principal Component Analysis (PCA), a manual iteration of different combinations to features, and finally settling on the lasso regularization method. We then used these chosen characteristics to make a model that can predict the future number of cases of COVID-19, and made visualizations to see our results. The data we used for our analysis was gathered from the Yu Group covid-19-severity prediction datasets containing information on COVID-19 cases/deaths, demographics, health resource availability, health risk factors, and social distancing which was useful for finding different features present in each county to use in our

model. We also made use of the JHU CSSE COVID-19 dataset containing a time series table of COVID-19 cases and deaths from 1/22/20 to 4/18/20 to help train our model.

II. DATA EXPLORATION

Our first goal was to make predictions on deaths and cases at a state-level, and this required aggregating the data in the "abridged_counties" dataset to get representative values of the columns for each state. Different columns, however, required different methods of aggregation. For instance, it wouldn't make sense to add the population densities of each county in a state to estimate the population density of the state. For this reason, we grouped each of the columns in one of 3 aggregation methods: mean, sum, and a weighted sum based on the percentage of state population made up by a given county. Then, we cleaned each of the columns by replacing NaN values with the column mean, and we used the grouping from earlier to group the counties by state and aggregate the columns appropriately. Next, we moved onto the "covid19deaths" dataframe and grouped the data by state to obtain values for the quantities we were trying to predict (number of cases and deaths per state). We replaced the state names in this dataframe with their abbreviations. This allowed us to join the cleaned and grouped "abridged_counties" dataset from earlier with the grouped "covid19deaths" frame. Subsequently, we created a dataframe that had both available features and values for the total number of COVID cases and deaths for each state. We were able to acquire the total number of cases and deaths by analyzing the latest data (given in the column of date 4/18/20).

The next step was shortlisting the group of all features to obtain a more relevant group of features to choose from. We began by removing all features that were clearly irrelevant such as the "Lat", "Long_", and "Country_Region". Then we created a correlation table to see which features are most associated with the number of deaths and cases respectively. Additionally, we used this visualization to aid us in selecting features with low collinearity, as this was something we looked to avoid when building our model.

The subsequent step in our data exploration was to use PCA to provide additional depth-of-understanding and to reduce the dimensionality of our dataset. We first applied a standard scaling of the data and then applied PCA using the *sklearn.decomposition* package (PCA is a form of Singular Value Decomposition). We started with 75 features in our original table and brought it down to 8 principal components in order to explain 95% of the variance in our data.

Using these observations, we aimed to base a model off the 3 largest principal components, ordered by their respective weights (singular values). The second heatmap (Figure 2) displays the contribution of each of our original 75 features to the 8 principal components. In our first model we select the 30 columns with the greatest contribution to the 3 most influential principal components. Since principal components are simply linear combinations of the different features, we hypothesized that the features with the greatest influence on the primary principal components explain most of the variance in the data. Thus, we assumed they would be the best features to form our regression model.

III. METHODS

Our first approach involved using the 30 features suggested by the aforementioned PCA analysis to predict the number of deaths for each state on 4/18/20 (the most recent total death count). We used a standard scaler to standardize the features and fit them using a Linear Regression Model. We utilized cross-validation and used a mean L1 loss to assess the model. We achieved a relatively high cross-validation loss of 600.9, which signified that our model lacked accuracy. We figured that the principal components involved a weighted combination of many correlated features (as visible in the long stretches of similarly colored cells in the first 3 rows in Figure 2). While the principal components themselves were orthogonal to each other, the underlying features which influenced each principal

component may have a high degree of collinearity. Thus, it was not advisable to base our model off the features which contributed most to the first three principal components.

Our next approach involved training a linear model by selecting features from a list of shortlisted features that passed a given threshold for correlation with the prediction variable. Using the mean absolute error of the model as a metric, we picked the combination of features that gave the least error. We discovered two problems with this approach. To begin with, the algorithm would take a long time to execute due to how quickly the number of possible feature combinations to iterate through increases as the number of shortlisted features increases. Secondly, after removing states that had NaN values for any of the features, we were left with data from only 30 states to train the model. As a result, we observed high cross validation errors of 81.3 and 125.8 for the number of deaths and cases prediction models respectively. Furthermore, we didn't incorporate regularization into training this model, indicating that it would perform even more poorly on the test data. We determined that we needed a model that was not only trained with more data points, but that also used a method of feature selection that didn't involve testing each possible combination of features.

Finally, we decided to try Lasso (Least Absolute Shrinkage and Selection Operator) Regression to try to select the best features from the dataset while parametrically balancing complexity to avoid overfitting. Before applying Lasso Regression to our data, we first removed columns from the dataset that were mostly NaN values like HPSAServedPop and HPSAUnservedPop because there was no logical way to fill these columns with numbers we can make up. We also dropped the ID columns for each county to prevent that from chosen as a feature. Next we filter out all the categorical variables because we cannot perform a one hot encoding as the categorical variables are the names of the county, state, and region and thus cannot be quantified. We also filled the NaN values for the "3-YrMortality" columns with 0's as NaN's in those columns meant that no people died in that age group hence no data for mortality rate. We also filled the NaN values for the "3-YrDiabetes" column with 0's as an NaN value indicates that nobody in that county had diabetes. By this point we have a cleaned feature table that we split into our training set and our test set using the train test split function.

Before using the data from our training sets in feature selection, we need to standardize the data using a standard scaler because the Lasso Regularizer of a linear model assumes that features are centered around zero and have variance in the same order. If a feature has a much larger variance relative to the other features it will prevent the model we are trying to create from learning from the other features as intended. We then use a select from model object on the features already scaled using our standard scaler to get the features selected by lasso regularization. Out of the 73 total features in the dataset, 29 were chosen through lasso regularization (Figure 9). Each feature was also assigned a different weight (Figure 10). Some features chosen were quite interesting, for example, the columns "dem_to_rep_ratio" and "Rural-UrbanContinuumCode2013" were chosen through the lasso model with weights of -5.24 and 0.14 respectively. This may be because those counties with high numbers of COVID-19 cases are usually urban areas with high populations that vote majority democrat. It is interesting how political affiliation can be used to predict the number of COVID-19 cases even though it may not have a direct affect on the likeliness of catching the virus. We also thought it was interesting how effective countywide policies to limit public gatherings were in stopping the spread of the virus. All of the columns relating to limiting public gatherings ("stay at home", ">50 gatherings", ">500 gatherings", "public schools", "restaurant dine-in", "entertainment/gym") were all chosen through the lasso method. This indicates the importance of early government action to curb large gatherings is vital to slowing the growth of COVID-19 cases. Such traits of the feature selection were different from that of our earlier model which only gave population and medical features, and we took this as a sign that our new model was more holistic. Before doing the feature

selection, we believed that the 2018 population estimate for each county would be chosen as one of the features because we believed that the number of cases would be directly proportional to the number of people in that county. However, it was not chosen through our lasso regularization meaning that other features had greater capability in predicting cases. This suggests that policies such as instituting a lockdown and features of the county itself such as whether it is in an urban or rural area has a greater effect on the number of cases than population alone.

To find our model's error, we initialized a linear regression to calculate cross validation error and fit our model over the features selected by lasso from the training set. We then calculated the test set error. Our model only had a cross validation error of 12.29 and a test error of 10.53, much more accurate than our previous attempts. To further increase the accuracy of our model, we graphed the cross validation error and test error over different values of the regularization parameter (Figure 4). The error was minimized at $C = 0.1$, so we kept the parameter at 0.1. Building on the relative success of our new model in predicting cases, we applied this model to predicting the number of deaths for each county and the number of cases for each state. We repeated the same process, and got an error of 14.65 for the test set for the number of deaths. However, with further analysis on the deaths prediction revealed that the error was only low due to many counties having zero deaths. Additional data on the number of deaths for each county up to now may improve our model's prediction when more counties have non-zero deaths. Training the data on a subset of the states with features table then testing it on the remaining data points for states resulted in a sizable error of 298.3. Since there are only slightly above 50 data points from the states table, we could not resolve the error in predicting the number of cases for each state using the features alone. We attempted training the model on the counties dataset then predicting the cases count for each state, but that resulted in errors in the tens of thousands. Data for other countries may also help train our model and strengthen our hypotheses. For example, many countries had different policies to put in place to prevent the spread of COVID-19. Our data was limited because each county had the same federal guidelines put in place so we cannot measure if the federal guidelines had any effect since every county followed them. Having other countries with different federal guidelines would have helped us test the benefit of each national policy.

IV. RESULTS

To visualize the accuracy of our best performing model of features selected through lasso regularization to predict the number of cases for each county, we created a line plot graphing our model's predictions for the number of cases for nine selected counties (chosen based on their large population sizes) against the actual case numbers for the time period between April 1st and April 18th (Figure 5). Based on the graph, our model's predictions generally follow the actual case numbers for each day. A major limitation of our model is that it requires the number of cases for the day before each prediction. This is why on the graph, both the actual cases and our model's predictions for April 1st are identical because we fed the actual number of cases for April 1st in our model, and kept feeding the actual case counts until April 17th. If our model was fed an initial number of cases of zero, it sometimes gave predictions of negative case counts which is a major problem for our model, but it performed better when there were already 10+ cases in that county.

V. CONCLUSION

When analyzing the data, we came across multiple ethical dilemmas. Our model predicts the most cases being concentrated around urban areas with high population densities. Is it fair that urban centers should receive better access to testing and medical care while rural areas are left neglected? Seeing how rural areas have hundreds of less resources such as hospitals or ICU beds was mind-boggling to say the least. It may even be worthwhile to predict the decrease in the number

of cases and deaths in rural areas if urban centers were to share their surplus of resources. Another finding of our model is that social distancing policies such as limiting public gatherings were very effective in limiting the growth of cases. Is it fair to those who cannot comply with those social distancing policies due to the nature of their work to restrict them from working, thus limiting their income? To make matters worse, blue collar workers who need their income the most bear the greatest burden of the quarantine sacrifice since most blue-collar jobs can't be done from home. While protecting people's health is most important during these trying times, we need to be considerate of individuals who are financially impacted by the quarantine as well. The data alone can only guide, but not answer these ethical concerns. That being said, the coronavirus pandemic is such that some sacrifices have to be made for the greater good to prevent the virus from spreading and causing greater damage.

VI. FUTURE WORK

In our primary model, we used the intrinsic features of the data, such as socioeconomic and related factors, to build, test, iterate, and visualize our cases and deaths predictor. Another approach to this involves mathematically modeling the growth of Covid-19 using past data and domain knowledge of pandemic models. We provide an introduction to possible future approaches by creating a rudimentary exponential growth model. By attempting to fit the pattern of deaths and cases and using epidemic models such as SIR, we may be able to increase model accuracy. Here, we provide a proof-of-concept introduction as to how we may be able to expand upon our model in the future.

We analyze the top few states, sorted by number of deaths, and select California, New York, Washington, and Michigan to build our models. As shown in the early (Figure 6) and overall growth graphs (Figure 7) , each of these 4 states has a completely different growth trajectory which makes these 4 states ideal for analysis. We selectively analyze the data after the 10th case in order to standardize our model for all 4 states, since we can pick an "initial value" for our exponential fit to be 10. Thus, we are able to reduce our analysis of growth down to the exponential growth factor term, instead of their initial values. From the length of the respective scatter plots in Figure 8, we see that California reached its 10th case quite early whereas Michigan is relatively new to the pandemic.

Using an exponential model with the equation

$$y(t) = 10(b^t)$$

$$y(0) = 10, \text{ standardized initial value}$$

$$t = \text{timestep in days since the 10th case}$$

$$b = \text{growth factor}$$

we find the best fit *b* parameter (the growth factor) using the *scipy.optimize curve_fit* module.

We find the best fit on the number of deaths in the first 3 weeks since the 10th case as a timeframe to model the "early" growth of COVID-19 in each state. In Figure 9, we see that California has a fairly low "early" growth factor which allowed it to contain the spread of Covid-19, whereas New York and Michigan had large early growth factors; Washington had a relatively large early growth factor as well. This may be explained by California's early response to the outbreak, being the first state in the nation to impose statewide shelter-in-place orders.

We then use the same exponential fit model equation to approximate the 4 states' overall growth factors (from their 10th case up to 4/18/20), not just for the first 3 weeks after their 10th case. In Figure 9, we see that all states were able to decrease their overall growth factors as time passed, indicating that the exponential model started to slow down. This indicates that a better model to use could be a logistic growth curve. This is further reflected in Figure 7, where the exponential model no longer seems to fit the data as well as it did earlier in Figure 6. This further indicates that an exponential model may not be the best for modeling the overall growth of the pandemic. For future work, we may instead consider using a logistic curve to better fit the data, as the possible effects of the shelter-in-place orders are seen.

Washington's implementation of strict shelter-in-place and social distancing measures may have led to its drastic decrease from early growth factor to overall growth factor. New York, which imposed its shelter-in-place orders after an initial outbreak, also saw a slowing down of growth rate.

Using rough benchmarks established by cursory analysis of growth factors of California, Washington, Michigan, and New York, we classified each state into one of 3 growth groups: Low, Medium, or High danger. We used the growth rates similar to that of California to indicate low danger, growth rates similar to that of Washington's early growth rate to indicate medium danger, and growth rates similar to that of New York's early growth rate to indicate high danger. A distribution of the growth factors of each state is shown in Figure 8.

VII. FIGURES AND TABLES

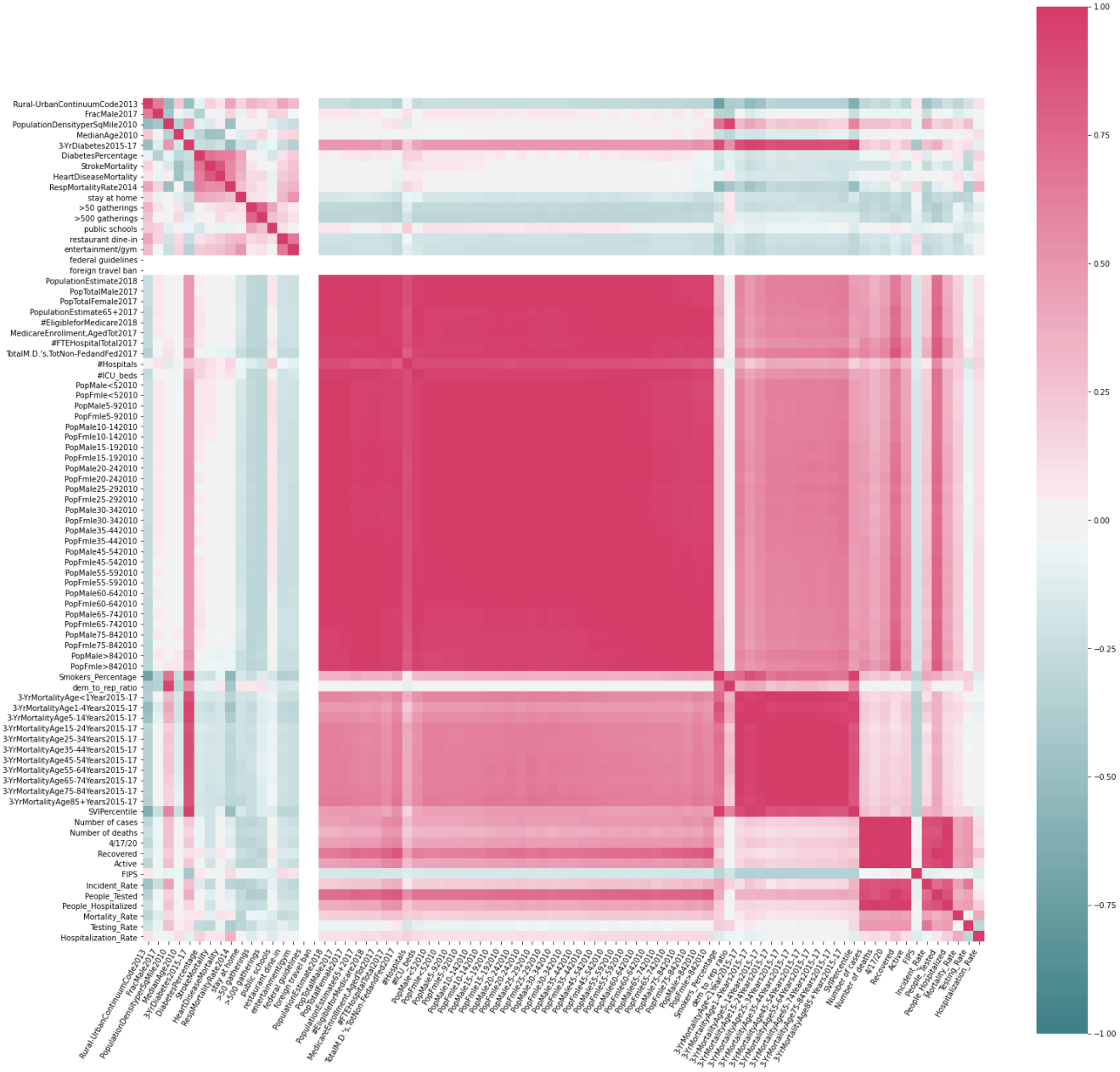

Figure 1. Correlation heatmap of all the features in the dataset for each state.

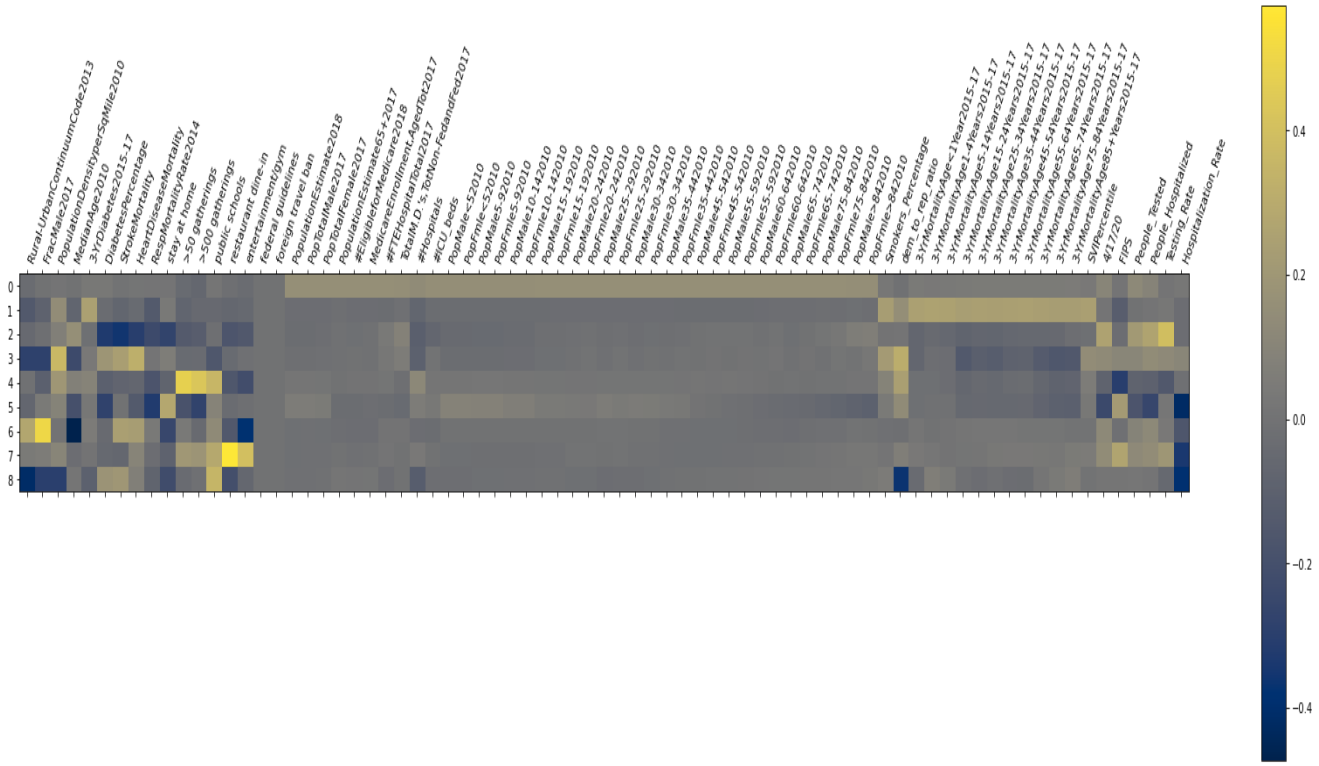

Figure 2. Heatmap of each of the features' contribution to the 8 principal components

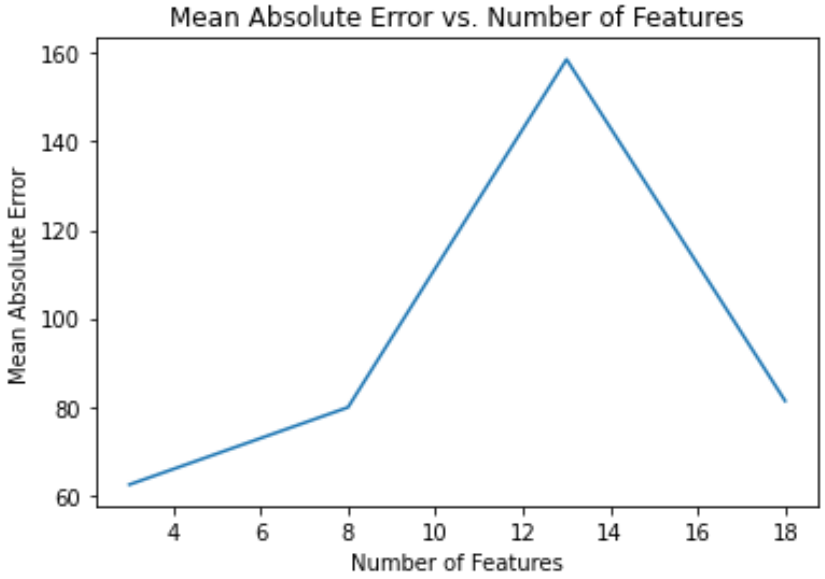

Figure 3. Graphing how the mean absolute error changes when the number of features are changed.

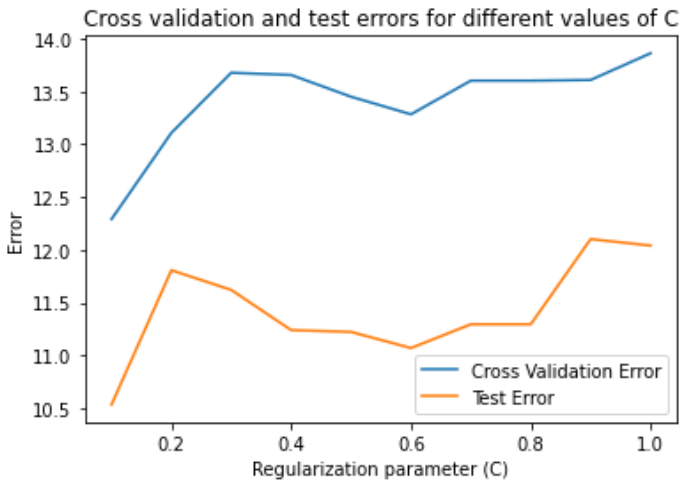

Figure 4. Graphing the cross validation and test errors for different regularization parameters when choosing features through lasso regularization.

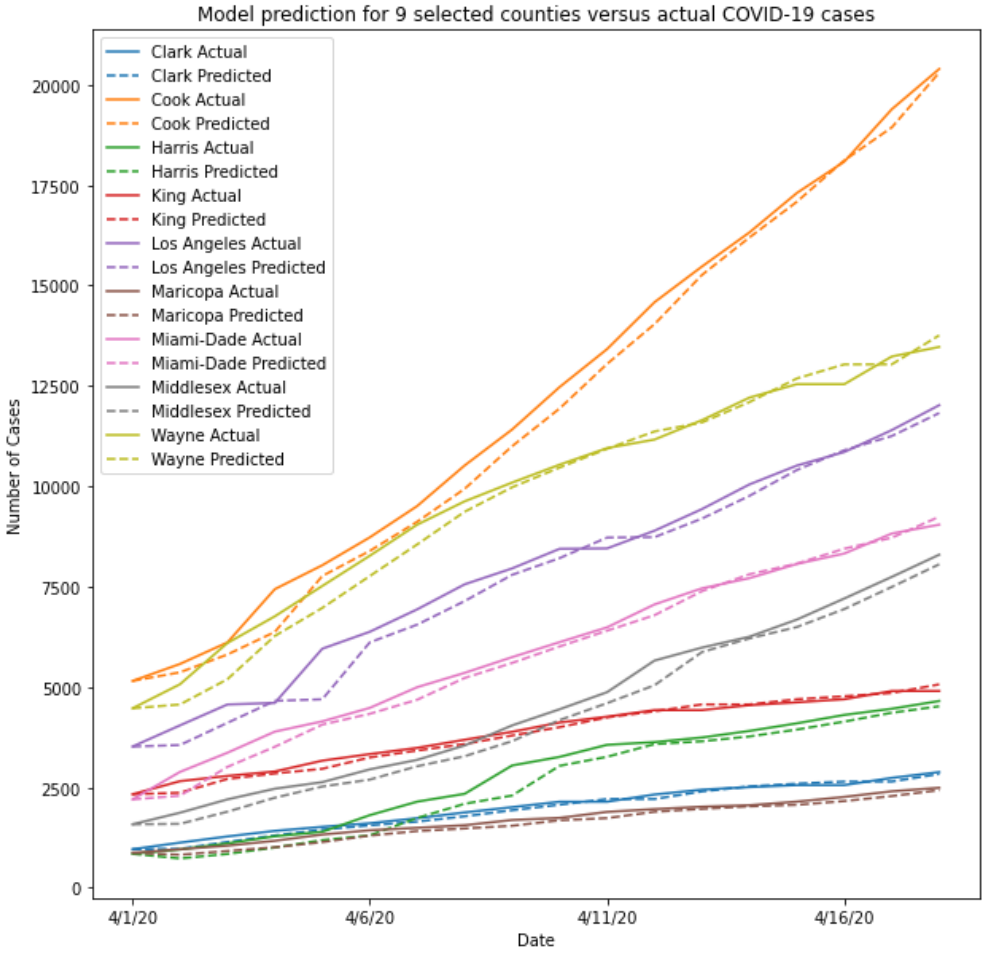

Figure 5. Model predictions and actual case numbers for the dates of 4/1/20 to 4/18/20 for 9 highly populated counties in the United States.

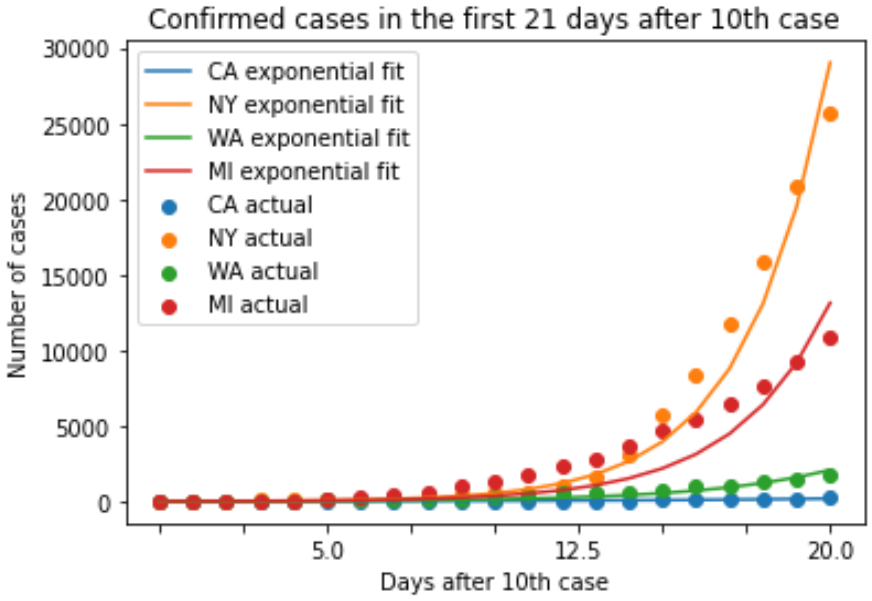

Figure 6. Graph of confirmed cases and exponential fits in the first 21 days after the 10th case for California, New York, Washington, and Michigan.

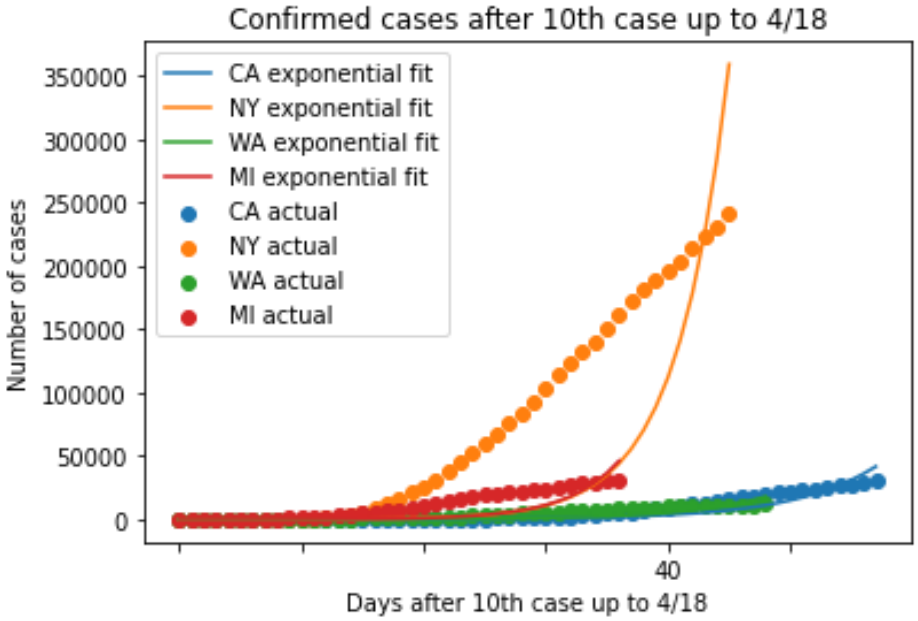

Figure 7. Graph of confirmed cases and exponential fits up to 4/18 after the 10th case for California, New York, Washington, Michigan.

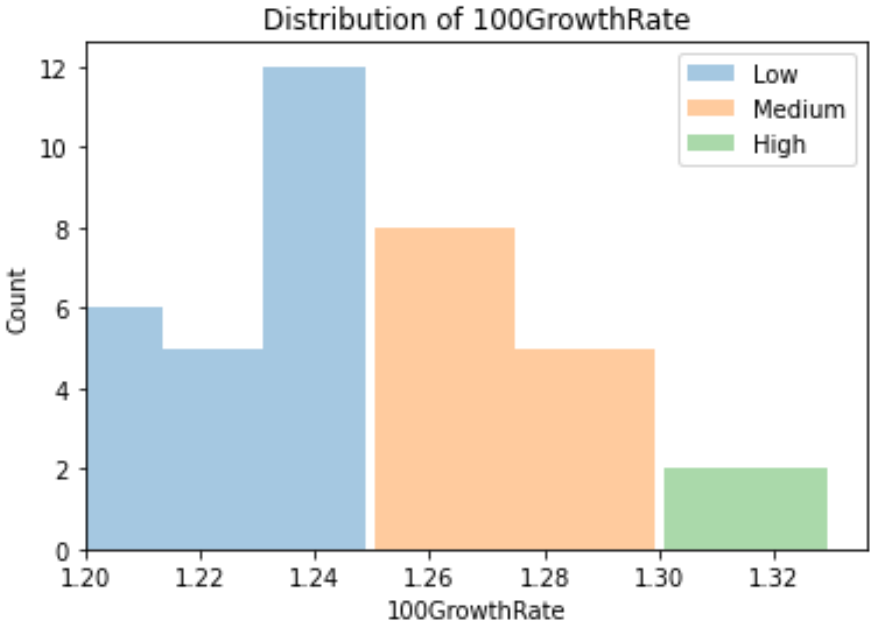

Figure 8. Exponential growth factor distribution for the states/provinces after the 100th case.

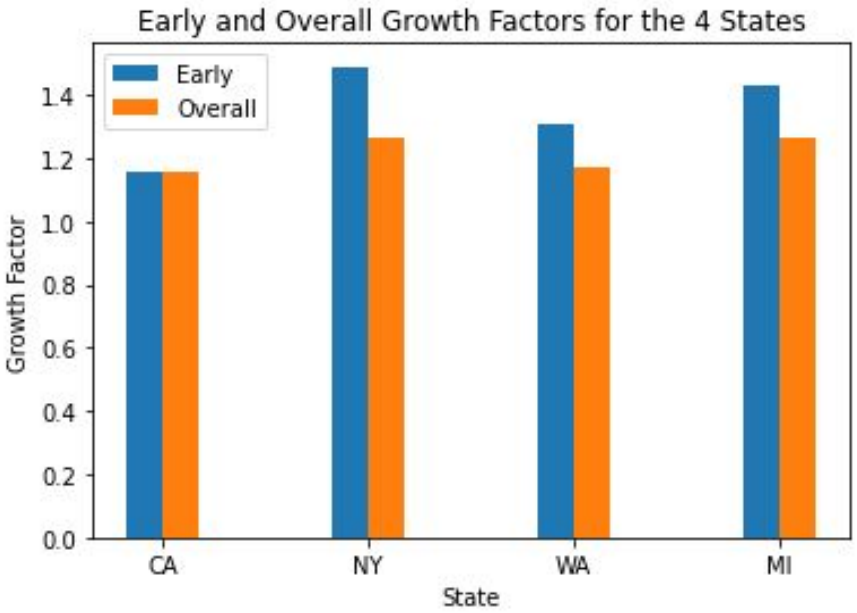

Figure 9. Comparison of "early" and overall exponential growth factors for California, New York, Washington, and Michigan

```
Selected Features from Lasso Regression: Index(['Rural-UrbanContinuumCode2013', 'FracMale2017',
       'PopulationDensityperSqMile2010', 'MedianAge2010', 'DiabetesPercentage',
       'HeartDiseaseMortality', 'StrokeMortality', 'Smokers_Percentage',
       'RespMortalityRate2014', '#FTEHospitalTotal2017',
       '#HospParticipatinginNetwork2017', '#Hospitals', 'dem_to_rep_ratio',
       'PopFmle15-192010', 'PopMale20-242010', '3-YrMortalityAge<1Year2015-17',
       '3-YrMortalityAge1-4Years2015-17', '3-YrMortalityAge15-24Years2015-17',
       '3-YrMortalityAge25-34Years2015-17',
       '3-YrMortalityAge35-44Years2015-17',
       '3-YrMortalityAge65-74Years2015-17', 'stay at home', '>50 gatherings',
       '>500 gatherings', 'public schools', 'restaurant dine-in',
       'entertainment/gym', 'SVIPercentile', '4/17/20'],
      dtype='object')
```

Figure 10. Chosen features using lasso regularization.

```
array([ 1.41338070e-01,  2.10991998e+01,  2.45794649e-02,  1.68580014e-01,
        1.62248434e-01, -8.99367013e-02,  3.24543003e-01,  6.23677701e-01,
        9.05073725e-02, -4.09414143e-04, -6.60724591e+00,  4.50218761e+00,
       -5.24323499e+00, -4.17291825e-03,  3.55771518e-03, -9.13236070e-01,
        1.41380385e+00,  1.04576222e+00, -4.87307923e-01,  1.77875269e-01,
       -2.68223452e-02, -3.23628621e-06, -1.75240326e-01,  1.75241431e-01,
        1.48791886e-01, -1.21624757e-01,  8.80942541e-07, -1.20034386e+00,
        1.04921409e+00])
```

Figure 11. Model weights for each feature chosen through lasso regularization.