Modeling COVID-19 Death Rate in the United States

Erin Lee, Joshua Kim

Feature Selection, Linear Regression, and Cross Validation on COVID-19 Data

Abstract

Using the most recent JHU CSSE COVID-19 dataset detailing COVID-19 confirmed numbers and deaths in the United States between January 22, 2020 to May 12, 2020 and the Yu Group's county data, we were able to explore different relations between the recent onset of COVID-19 and county health/demographic data. Based on the growing danger and alarming number of deaths in the US, we explored what factors most correlated with the death rate in counties by COVID-19 using sklearn.feature_selection's SelectKBest module. We then used cross-validation to select the best number of features that could reduce RMSE (Root Mean Square Error). While linear regression turned out to not be a very strong indicator of death rate, it still allowed us to observe key correlative factors.

Keywords: Feature Selection, Linear Regression, COVID-19, Death Rate, USA

Introduction

Coronavirus, specifically COVID-19, hit humanity where it least expected it in 2020. Affecting all layers of societal interaction, the containment and treatment of this new global pandemic were prioritized by numerous nations: leading to the lockdown of international communities. While certain countries were more effective at this stage than others, the United States has been a country that has been particularly hit hard by this new pandemic threatening human lives. Already, over 84,000 have lost lives in the three-and-a-half-month period since coronavirus was introduced to the United States, and many more are to come. Today, using the latest JHU CSSE COVID-19 dataset detailing COVID-19 confirmed numbers and deaths in the United States between January 22, 2020 to May 12, 2020 and the Yu Group's US county data, we plan to explore what social factors most correlate with the death rate in the United States by COVID-19.

Exploratory Data Analysis & Data Visualizations

We first began by examining the above datasets for possibly interesting correlations or trends in the data that could possibly be helpful in feature selection or engineering. While creating s multitude of different visualizations we came to notice several interesting features that we hadn't known of before. One of these was the "HPSA Shortage" feature that represented the "number of full-time equivalent practitioners needed in the Health Professional Shortage Area (HPSA) so that it [would] achieve the population to practitioner target ratio" (Yu-Group). While we weren't able to make a meaningful visualization based on this feature alone due to many counties missing this data, it did point us in a direction to further examine medical resource availability in the form of number of ICU beds and hospitals in counties and their relationship with population and COVID-19 death rate (Visible in Figure 2, 4, and 5). Another interesting

feature that pointed us in a specific direction was the SVI Percentile (or Social Vulnerability Index Percentile). The SVI Percentile represented the county's social vulnerability and the greater the percentage, the more vulnerable they were to be against pandemics like COVID-19. Since a higher SVI Percentile essentially meant weaker immunity against COVID-19 we initially believed the SVI Percentile to have a close relationship with death rate by COVID-19. Contrary to our beliefs though, as we can see in Figure 2, there wasn't much of a significant relationship between these two factors. Similarly, while we believed that a higher smoking percentage, possibly indicative of weaker lung health, would correlate with weaker immunity against COVID-19 and raise death rates, this correlation was also close to nonexistent as visible in Figure 1. Finally, we looked at the geospatial density distribution of death rates of COVID-19 and noted COVID-19 death rates being particularly higher near eastern coastal sections of the US (Figure 6).

Data Cleaning & Transformations

In order to compare death rate by COVID-19 in counties in the United States to the several features made available in the Yu-Group county data file, we first had to clean the confirmed and death timeseries files so that they could be joined to the county data file. We did this by first summing all of the confirmed cases and death cases by county and joining the two files using the "FIPS" index. We then created a new column called "death rate" that took the number of total deaths by COVID-19 to date in the county divided by the total number of confirmed COVID-19 cases in the county. Having done this, we joined this combined timeseries data file to the county data file using the FIPS index (countyFIPS in the county data file) again to create a unified file named comp_table.

Since we aimed to ultimately find relations involving the death rate in counties, we first filled all null-values in the death rate column with 0s using the fillna() method and filtered the comp_table for only cases of death rate > 0 and total confirmed number of cases > 1000. This was done so that we could filter for outliers and counties that had only recently been introduced to the virus. We then created some of our own features using our findings from the EDA step like number of ICU beds per person, percentage of population enrolled in Medicare, and elderly ratio (using the 2010 census, total number of elderly > 65 years of age divided by total population). However, at this point we faced several challenges with our data in both its vastness and limits. Especially true for the county data file, while the number of features it held were numerous, there were multiple columns with numerous null values that we had to fix before being able to use. While we first filtered for columns with entries of 60% or more null values, we then faced the ethical dilemma of having to choose what method to deal with the remaining null value entries. It was possible to delete all rows with missing entries, but believing that it would further corrupt our data, we instead decided to fill all null entries using the median value of columns. Several limitations in the data like the mismatch of year (2010, 2018, and 2020 all existing in a single file) and lack of more detailed situations of coronavirus patients did bother our understanding of coronavirus related death rate and stall us, but we were able to instead focus more on demographics, medical risk factors, and social mobility within counties.

Methods & Justification

Before we selected our features and tested for linear regression, we first began by splitting our dataset into training and test sets (9:1 ratio) and standardizing the x_train and x_test sets. By standardizing, the features within the datasets gain similar weight and help better see correlations. We then used the SelectKBest and f_regression module from the

sklearn.feature_selection library to help in our search for the five best linear indicators of death rate in counties by COVID-19. We decided to use linear regression to test for this relationship since we wanted to be able to weigh each feature approximately equally in our search for the most correlative linear factors relating to death rate in counties by COVID-19. Had we used Lasso or Ridge Regression, while we may have been more successful in creating a more accurate model, being able to distinguish the features that most correlate with death rate by COVID-19 may have been difficult. Thus, even if we sacrificed accuracy of our model, we believed it necessary to use a model that could better illustrate feature importance. The use of the f_regression and SelectKBest module also helped us to easily and efficiently obtain the five and ten most important features to univariate linear regression (the SelectKBest module selects the highest scoring features based on the scoring module, f_regression). We finished by analyzing the R-squared value, training & test RMSE, and cross validation RMSE.

Analysis & Conclusion

Based on our results, we found that the five most features indicative of a linear relationship to death rates by COVID-19 were the features: Median Age 2010, Heart Disease Mortality, Stay at Home, Elderly Ratio, and Medicare Percentage. These five factors showed a R-squared value of .0912, Training RMSE of .0319, Test RMSE of .0326, and cross validation RMSE of .0320. The low R-squared value indicative of a low linear relation between these variables and death rate by COVID-19 show that the relationship between these two are most likely not linear in nature. Doing a similar analysis for the ten best factors, we saw a small increase in R-squared (.0977), Test RMSE, and cross validation RMSE while training RMSE showed a small drop as is natural with the increase in number of features. Finally, doing an

analysis of the cross validation RMSE for number of features used in training set revealed five features to be the most optimal in reducing cross validation RMSE (Figure Seven).

Limitations and Discussion

As a result of our decision to use a linear regression model, there were several limitations made to the analysis we did. While we assumed that there were would be some linear relationships between death rate by COVID-19 and the features given to us, our analysis proved this to be quite wrong with our weak R-squared value. If we had used a Lasso or Ridge model we may have been able to achieve a lower cross validation error as well, though it may have strayed away from our research question. Furthermore, our model of the death rate didn't take into account rate as a derivative and rather simply divided the number of deaths by the number of people confirmed with COVID-19. This assumption may have made a large difference in our understanding of the relationship as a linear one and the results as well. If we were given more data about the patients who died of coronavirus – average number of days until death, race, average age, etc… – it may have proved extremely helpful in understanding how to make better features that would have better correlated with death rate by COVID-19. Finally, studying this issue, there may be ethical concerns arising involving a respect for privacy and anonymity. To better understand this virus and its effect on patients, it is inevitable that researchers better understand patients and their surroundings and behavior. However, many times where the line is drawn may be gray and difficult to maintain. We believe that the only way to address these concerns is through a combination of informed consent and continuous reminders and communication between patients and researchers regarding boundaries involving privacy and anonymity. After all, patients too are just a little bit unluckier people.

Figures

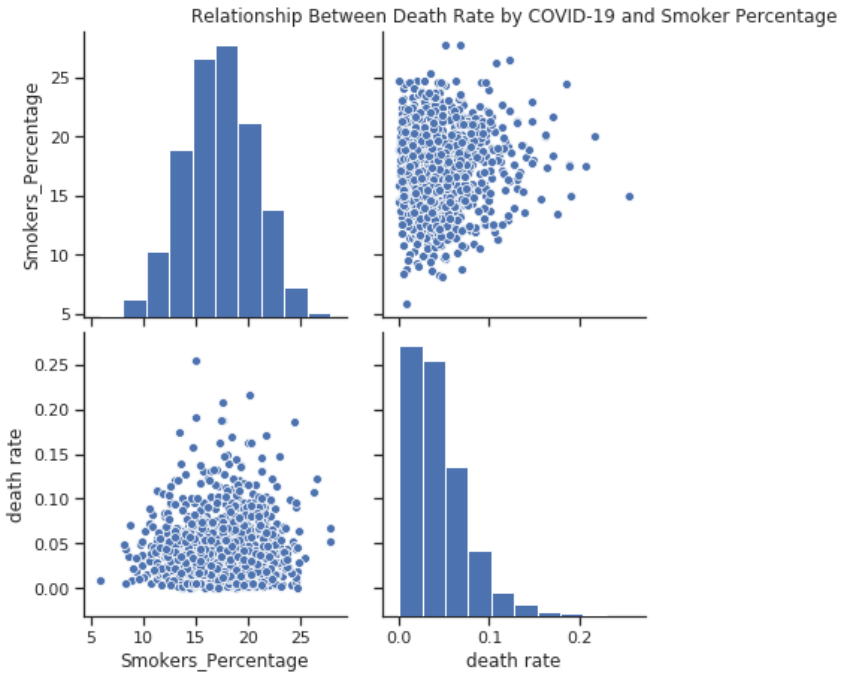

Figure 1. Relationship Between Death Rate by COVID-19 and Smoker Percentage in Counties

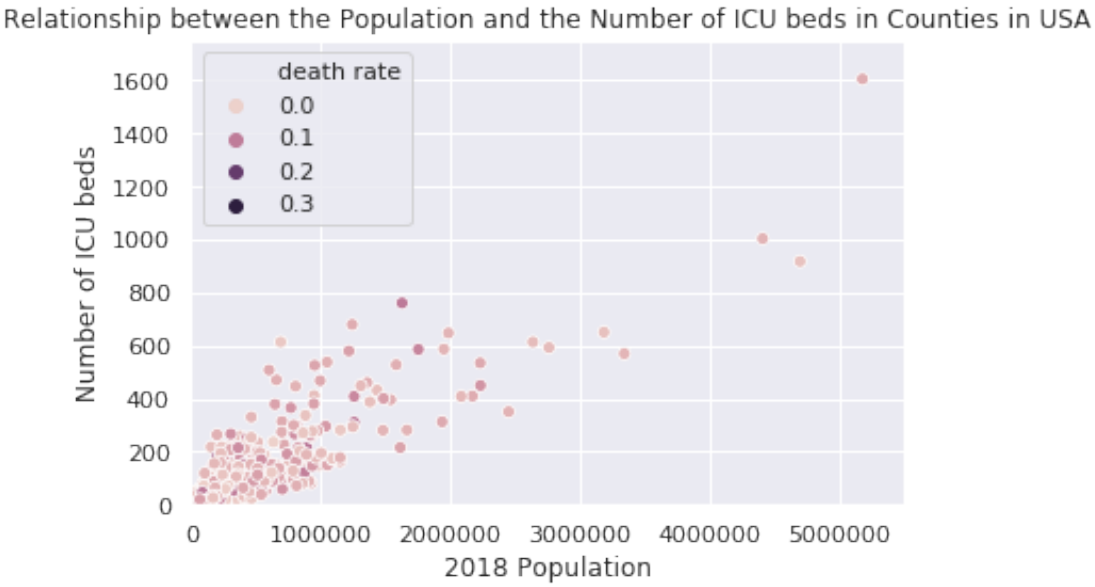

Figure 2. Relationship Between 2018 Population in Counties and Number of ICU Beds in Counties

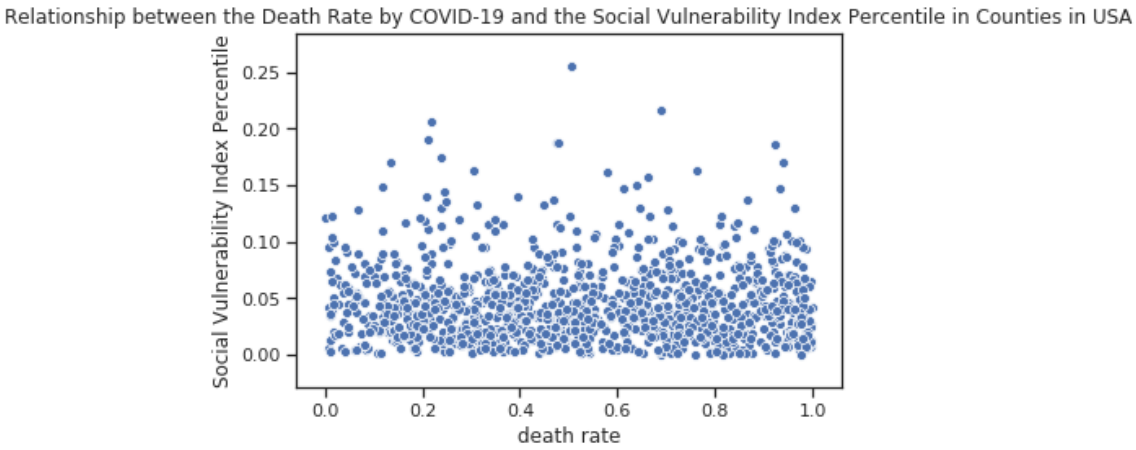

Figure 3. Relationship Between Death Rate by COVID-19 and the Social Vulnerability Index Percentile by Counties

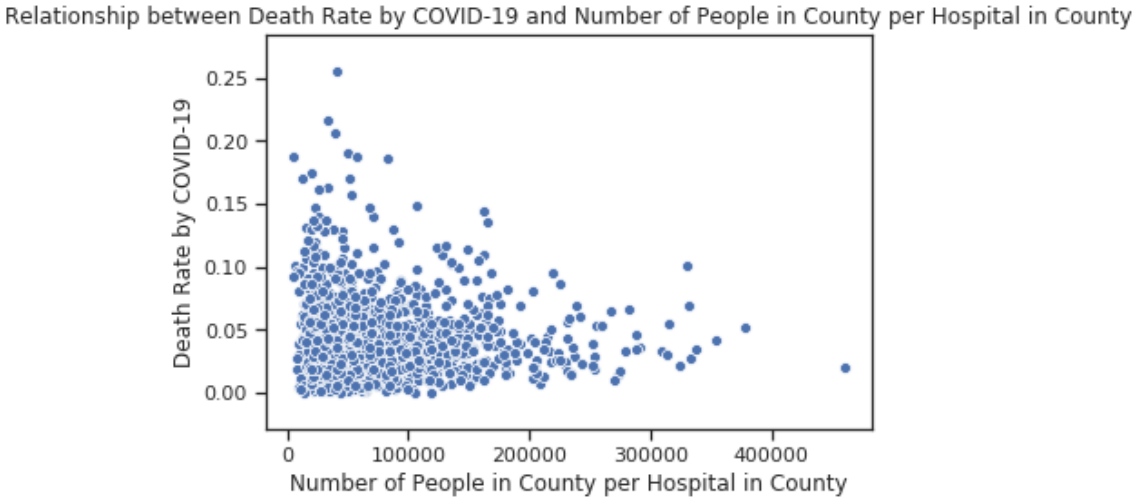

Figure 4. Relationship Between Death Rate by COVID-19 and Number of People in County per Hospital

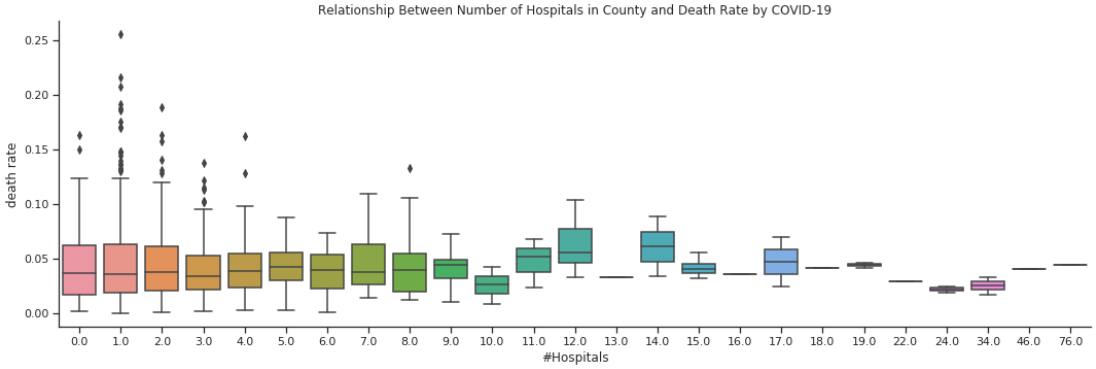

Figure 5. Relationship Between Number of Hospitals in County and Death Rate by COVID-19

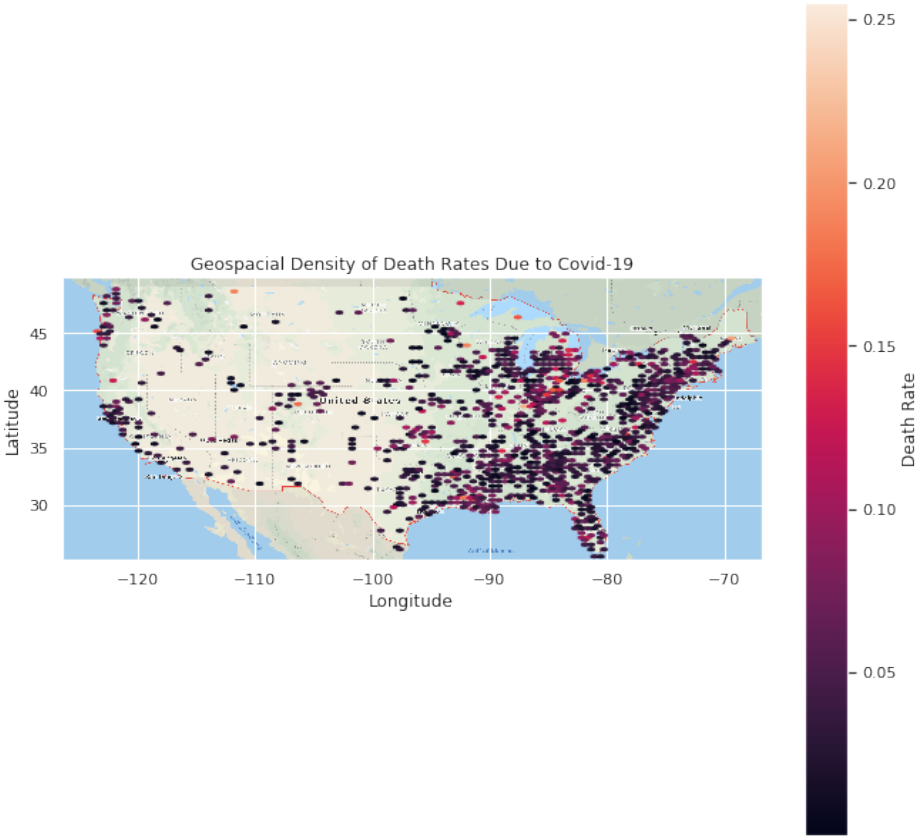

Figure 6. Geospatial Density of Death Rates Due to COVID-19 in US

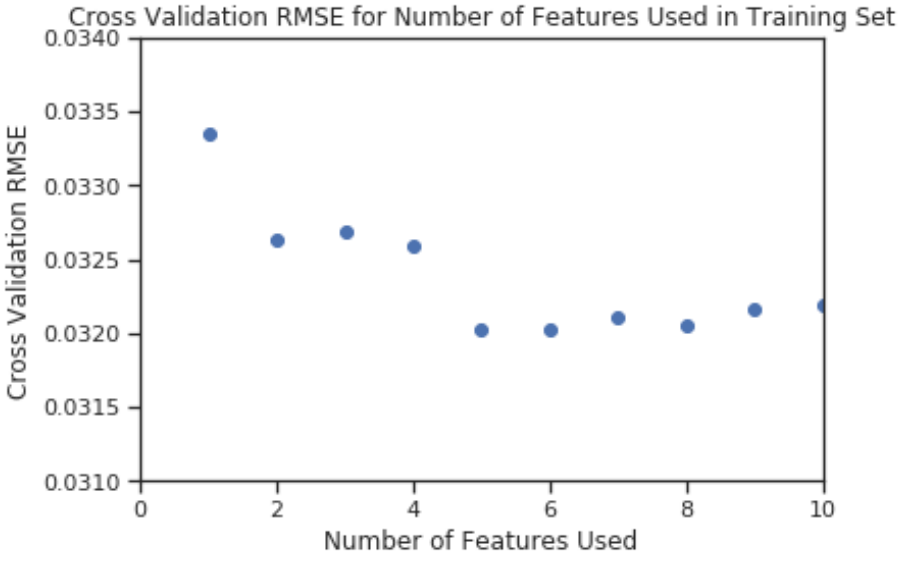

Figure 7. Cross Validation RMSE for Number of Features Used in Training Set