Predicting the Number of COVID-19 Related Deaths per county in the US

Authors: Eric Cheng, Kevin Liu, Arbaaz Muslim

**Abstract**

Since emerging in late 2019, the novel coronavirus, now referred as COVID19, has caused nearly 300,000 deaths to date. Our goal was to predict the number of deaths by county, through analysis of over 3000 counties through the United States. We wanted to see which demographic factors lead to deaths and model the number of cumulative deaths in each region. We utilized Lasso Regression and Random Forrest in our models and referred to R^2 and RMSE as our benchmark for accuracy in tuning our models through cross validation. Upon completion, we saw that population density as the driving factor between spread of COVID19 within counties, with stay at home orders and limits of gatherings as contributing factors. The number of hospitals and doctors also played a limited, but non-trivial role as well. Based on our observations, we conclude that population control plays a significant role in preventing the number of COVID19 deaths.

**Introduction**

In this project, we are interested in answering the question of how the actions or demographics of a certain county or city play into the number of COVID-19 related deaths. Across the United States, some counties say a staggering number of COVID-19 deaths while other counties saw only a low number of deaths. We wanted to see if the some of the demographic factors, such as population, # of hospitals, or the obesity rate of a certain county were helpful characteristics in predicting the number of COVID-19 related deaths in a county. As such our question is: *How do we model the number of deaths from the demographics of a particular county?*

**Description of Data & Data Cleaning**

We mainly utilize two data frames of COVID-19 related information.

1. The timeseries of COVID-19 related deaths starting from 1/22/20. This was from JHU and we utilized updated data until May 7th, 2020.
2. The county level data, which contains different characteristics of counties (e.g. population, mortality rates, dates of important COVID-19 orders)

We also created our own data frame:

1. A data frame called supplemental\_data  consisting of States and State abbreviations (California -> CA) to help fill in NAN values in States.

The columns that we used from timeseries (1):

* Admin2:County Name
* Province\_State:Province, state or dependency name
* Columns representing number of cumulative deaths from 1/22/20 to Present

The columns that we used from county level data (2):

* POP\_LATITUDE:
* POP\_LONGITUDE:
* PopulationEstimate2018:
* PopTotalMale2017:
* PopTotalFemale2017:
* FracMale2017:
* PopulationEstimate65+2017:
* PopulationDensityperSqMile2010:
* CensusPopulation2010:
* MedianAge2010:
* #EligibleforMedicare2018:
* MedicareEnrollment,AgedTot2017:
* DiabetesPercentage:
* HeartDiseaseMortality:
* StrokeMortality
* Smokers\_Percentage
* RespMortalityRate2014
* #FTEHospitalTotal2017
* TotalM.D.'s,TotNon-FedandFed2017
* #Hospitals
* #ICU\_beds:
* Dem\_to\_rep\_ratio
* stay at home:
* >50 gatherings:
* >500 gatherings:
* public schools:
* restaurant dine-in:
* entertainment/gym:
* SVIPercentile:

The columns that we used from supplemental data (3):

* State
* StateName

We first used five properties of exploratory data analysis: structure, granularity, scope, temporality, and faithfulness to clean the two dataframes.

When we saw NAN values, we first investigated which rows contained those NAN values and looked for any specific trends. For example, a certain number of states had NAN values for stay at home orders and our investigation found that those states never had a stay at home order, so we left them as is (or assigned an arbitrary date far into the future). Another example was when we saw that many counties in Alaska had NAN values for the proportion of democratic to republican votes. Domain analysis revealed that those rural counties did not have a political preference, so we simply replaced those NAN values with a “1” indicating an even split. And in other cases, we simply dropped those rows with NAN values (for exp: mortality rate) that were both hard to calculate and trivial in quantity (there were only a few rows with these values).

**Methodology: Analyzing Correlations, Plotting Visualizations, and Model Fitting**

Next, in order to see which parameters affected the number of deaths per county the most or observe any peculiar patterns in the data, we decided to create a few visualizations to analyze the correlations between certain columns with the current number of deaths in our death\_and\_demo dataframe. We began by doing some exploratory data analysis and domain research and started off with a correlation map to see which features would be interesting to look at. In our correlation map (Figure 1), we see that the population density, the number of doctors and hospitals as well as stay at home orders were interesting characteristics to look at. We are also fully aware that many characteristics may have been collinear.

A screenshot of a cell phone

Description automatically generated

Figure 1: Correlation Map (partial, full map in notebook)

We saw that population density was positively correlated  and decided to investigate this with a scatter plot (Figure 2), which proved our initial observations.

A screenshot of a social media post

Description automatically generated

Figure 2: Scatterplot of Deaths and Population Density.

With this idea in mind, we proceed to create our model and perform feature selection. Because we had a large number of features, we fit our pipeline using a LASSO model. We iterate over 500 alpha values in order to find the one with the lowest cross-validation error, with the

𝑅^2 value as our metric. We choose this as our metric because in this phase, we are focusing on creating a model that is interpretable. The LASSO model also has high values for its tolerance and maximum iterations parameters. In an earlier iteration of this notebook, we ran the pipeline without these arguments and got numerous Convergence Warnings. In addition to these warnings, our "best" model had a negative  R^2  value. We added these arguments in order to resolve these issues. After running our model, we saw  that the best performing alpha value is about 307, as illustrated in Figure 3.

A screenshot of a map

Description automatically generated

Figure 3(CV score vs Alpha Values)

We then fit our model to our data and were able to determine that population density, stay at home orders, and limits of gatherings of over 50 people were key factors in determining the deaths in a county. For full details of the fitting process and coefficient values, please refer to the notebook.

Now, we proceeded to begin the regression phase of our project, where we used  the data in order to predict the death count in a county. Instead of the R^2 value, our metric is now the root mean squared error, since its units are more interpretable. Our development of a regression model is four parts: a decision tree regressor, a random forest regressor, and 2 SVMs, each with a different kernel. For the development our models, please refer to our notebook as it has been omitted in this report for brevity.

To summarize, we can visualize the RMSE for each model. Figure 4.1 is with SVR Regressor (Linear) omitted due to skewing of the chart.

A screenshot of a cell phone

Description automatically generated

Figure 4 (Model comparison)

Of the remaining models, the random forest regressor seems to follow the classic pattern of overfitting, with a lower train error and a higher cross validation error. Since we know that the random forest model can benefit from some form of regularization, we continue onwards with it. When we read through the documentation for the random forest regressor, two arguments to the model caught our attention: the max\_depth and n\_estimators arguments. We realized that testing smaller values for either argument would serve as a proxy for a regularization term and would reduce overfitting. We chose to tune the n\_estimators argument.

Figure 5 gives us a nice visualization of how the model performs as we tune the n\_estimators argument.

A close up of a map

Description automatically generated

Figure 5 (N estimator graph)

As with our first model, we fit our model with the data and see that population density had a key role in our model, confirming our initial finding.

**Summary of Results**

We saw that population density was a significant factor in our model when predicting the number of deaths within a county. We were surprised by this as we thought that population would be a dominating factor (with one being the multiple of the other). Rather, the density of population mattered much more. Our lasso model revealed that stay at home orders, and limits of gatherings of 50 or more were also important factors in the model. Our random forest regression model also produced similar results noting that population density was the dominating factor, and the number of hospitals and doctors also playing a limited role in determining the number of deaths. In our analysis, the RMSE for Random Forest Regressor was between 150 and 200 for our train set and 300 to 350 in our CV. In our test set, we obtain a value between 150 and 200.

**Discussion**

We were surprised to see that that population density plays such a dominating role in the number of deaths in a county. While we expected that a higher population naturally would exhibit more deaths, our models relied on population density as the dominating factor.

We had expected the median age to also play a role as early research shows that COVID19 affects the elderly population more often, but upon investigation realized that in the vast majority of counties, the median age was well below 65 which did not help our analysis. We also realized that many factors may be colinear, and because our domain research was limited, could not pinpoint each to each individual factor.

One challenge that we faced was undoubtedly fitting our model. As documented in our notebook, we ran into convergence errors and the number of iterations to run. We also took time to research into different arguments when creating out model (Lasso and Random Forrest) that were unfamiliar. A lot of debugging and ensuring the right syntax was done through trial and error, as the libraries used were fairly new to us.

A major limitation in our model is that we did not account for collinearity as our domain research in infectious diseases is not strong. It remains difficult to determine which characteristics were colinear to each other, and if we had more time, we would perform more analysis and research to solve this problem.

An ethical dilemma was deciding whether or not to include certain minor outlying United States  islands  and territories that were in our dataset. While we were aware that excluding these territories may bias our model against smaller populations or underrepresented minorities, including the data would also pose a logistical challenge as many of those territories and islands had an abnormal population, incomplete data, and obscure state abbreviations. In addition, we were concerned that including these territories, many of which were outliers, would skew our model which ultimately led us to make the difficult choice of discarding those territories. Our driving decision was to minimize bias and discarding those territories made the most sense.

Additional data that would have strengthened our hypothesis would be mortality rate for each county. We had the mortality rate for each state but had originally wanted to predict the mortality rate for each county but did not have this information. Additional factors that may have been helpful, according to our domain research, would be financial information for each county, as we know that typical diseases disproportionately affect lower income communities, as well as education level. These factors would have strengthened our hypothesis.

Originally, we had also wanted to create a classifier to calculate whether a person who got COVID19 would die or recover. However, we realized that we did not have data per person, and only at the county level.

Throughout our analysis ethical concerns were minimal. All contributors involved in this analysis did not have any immediate conflict of interests and did not contract COVID-19 so we were not biased by our own experiences. Further, we were glad that data provided was normalized so we did not access any private or otherwise secure information. The only ethical concern was that because this was a sensitive topic, we tried to present our finds and analysis in a way that showed empathy while preserving our objectiveness. We understand that these topics may bring grief to some and avoid using any inflammatory language.

**References**

*Sci-Kit Learn API Reference*: <https://scikit-learn.org/stable/modules/classes.html>

*Data from JHU:* <https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv>

*Data from Yu Group:* <https://github.com/Yu-Group/covid19-severity-prediction/blob/master/data/county_data_abridged.csv>:

*Geography of COVID19:* <https://www.citylab.com/equity/2020/04/coronavirus-spread-map-city-urban-density-suburbs-rural-data/609394/>