The Future of COVID-19: how many related deaths will occur and determining risk?

Jillian Criscuolo and Quin Etnyre

Abstract

In this analysis we seek to answer two very important questions in the world of COVID-19. First, we explore the *states418* and *abridged_counties* to identify if a relationship exists between the total number of confirmed cases and total number of people tested for COVID-19. Additionally, we sought to find a relationship between the deaths and total COVID-19 cases for those tested and those confirmed positive. Through running logistic regression to predict the amount of future deaths on the states with over 15,000 confirmed cases, we sought to address a new question: can we predict whether a county is high risk or moderate risk, depending on the amount of people that have been tested and the total number of confirmed positive cases in that specific state. The high-risk states are defined as states with more than 15,000 confirmed cases. Throughout the entire analysis New York is identified as an outlier and was completely removed from the initial question of how many future deaths there will be in high-risk states, but remained as part of the data for the second models that categorized counties as low or moderate risk based on the metrics from their state.

Analysis

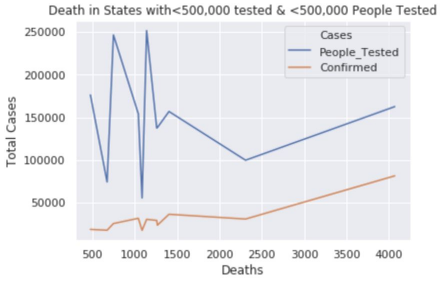
Description of Data and Methods

We performed EDA to look at the *states418* dataset that has rows representing the April 18th entry for each state on the number of confirmed cases, deaths, recovered, active, the testing rate, people hospitalized, etc. We identified that Wyoming was missing from this dataset. The *abridged_counties* dataset was also examined and contained different counties' information on hospital-level data, demographics, health risk factors, and social distancing factors. The *nullcounties* dataframe contains the counties whose state's are excluded from the *abridged_counties* dataframe. To combat this, an external .csv file was added that contained all 50 US states, including District of Columbia. This .csv file was added in as *states* and was merged with *abridged_counties* to create *addedstate* which filled all NaN values that were present in *nullcounties*. We also filled all NaN values in the smoking population and total with the column mean as it made more sense to generalize it to all counties/states as opposed to filling them with 0, counting their populations as zero due to lack of data for 11 counties.

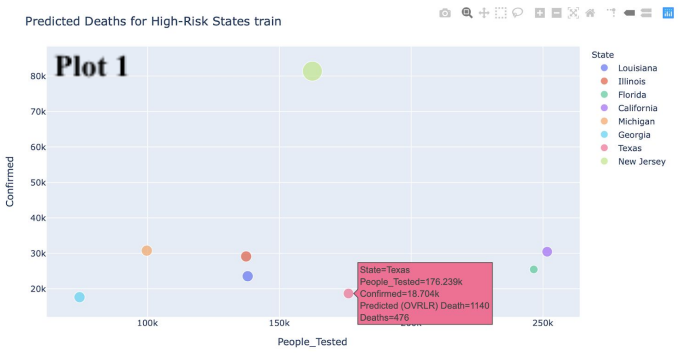
For the first part of our analysis that sought to predict the number of deaths based on confirmed cases and people tested we decided to merge *abridged_counties* and *states418* and explore, for each state, the number of people tested, confirmed cases, deaths, and population as these were the most consistent columns. We assumed the smoking percentage would be useful as COVID-19 is a respiratory virus, but it turned out to be ineffective because the amount of the population for each state that smoked was negligible compared to the amount of COVID-19 cases. Additionally, there was not an accurate way to conclude whether those who had tested positive were among the smoking population in that state and it turned out Louisiana was the only state with more than 15,000 confirmed cases that was in the top five states for smoking.

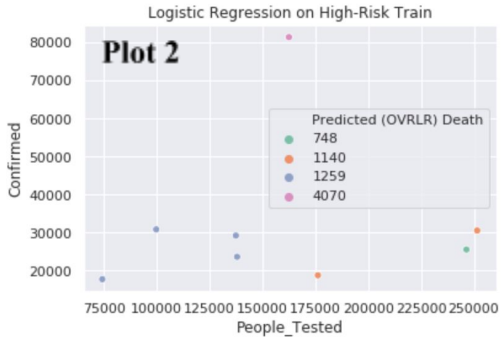
A challenge we found with the data was when merging the two tables and deciding whether or not we wanted to preserve the county FIPS, but soon realized it would be better to observe the states as a whole as information was available on the mortality and number of cases by the state–not county. We also attempted to explore the active percentage of patients positive with COVID-19 but saw that it would not help us achieve our goal of predicting

deaths in the future as there was a clearer correlation between people tested and people confirmed. We decided to limit the final data set to use for prediction to just states with over 15,000 cases as the rest of the states became negligible when skewed with high-risk states. New York became a special case as it had almost half a million people tested, becoming an outlier on this dataset. We decided to remove New York in the *confirmed_testny* data frame and immediately a graph that was much more representative of the high-risk states defined as more than 15,000 positive cases and less than 500,000 people tested.



When splitting the data an area that caused confusion at first was performing logistics regression on nonbinary features. However, we realized it would be better to use logistic rather than linear regression because of the rapid growth in the beginning and we could make predictions based on preexisting data where logistic tapers off for pandemics and could predict deaths by categorizing the states in plot 1 and the number of deaths in plot 2 for the training data.

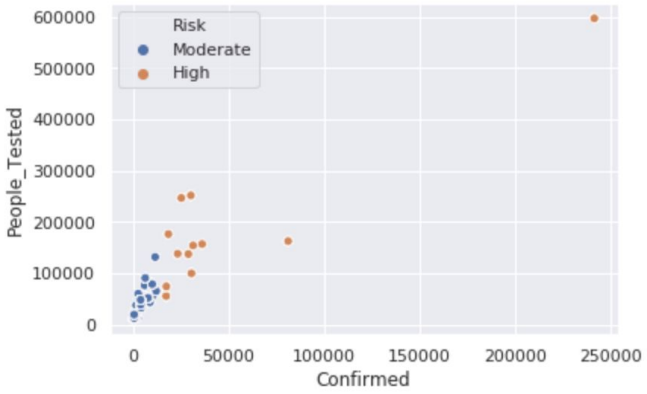


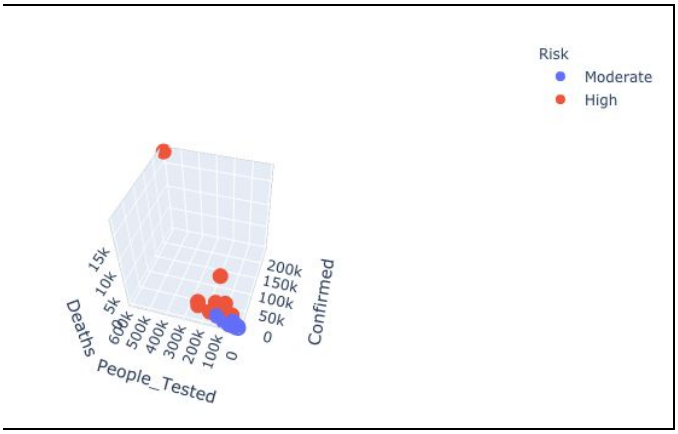


Logistic regression proved to not be the best form of prediction modeling when dealing with a very small dataset not made up of purely categorical data. Visually it seemed pretty accurate; however, when using the logistic regression .score function it proved to be extremely over fit with training accuracy around 0.5 and testing around 1.0.

	State	People_Tested	Confirmed	Deaths	Predicted (OVRLR) Death
18	Louisiana	137999.0	23580	1267	1259
13	Illinois	137404.0	29160	1259	1259
9	Florida	246527.0	25492	748	748
4	California	251614.0	30491	1140	1140
22	Michigan	99727.0	30791	2308	1259
10	Georgia	74208.0	17669	673	1259
43	Texas	176239.0	18704	476	1140
30	New Jersey	162536.0	81420	4070	4070

We instead decided to add the state name from the *states* dataframe to the *abridged_counties* dataset and categorize all counties dependent as high or moderate risk dependent on their state and stored this in *combinedcounties*. High risk is categorized as any state with more than 15,000 confirmed cases and moderate risk is any state with less than 15,000 cases. Unlike the previous analysis we decided to use all the counties to predict whether a state was low or high risk. The features we decided to use were total number of people tested and total confirmed cases. This created a more positive correlation when looking at all the counties. We also created a 3d plot to add in a third dimension of the amount of deaths in each county based on total deaths in the state.

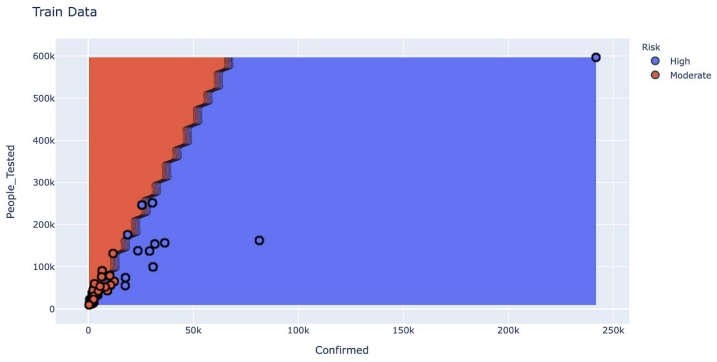


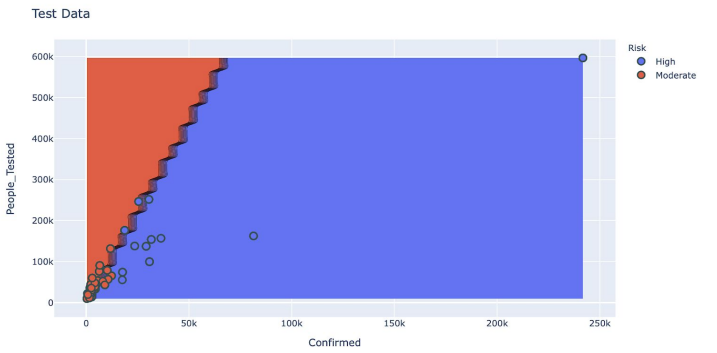


combinedcounties was split into train and test subsets and fit the data using one-vs-rest logistic regression that we saw worked best with multiclass classification. It builds one binary logistic regression classifier for each N classes– in the *combindedcounties* dataframe N=2. Again, visually the Logistic Regression using the OVRLR model on the train set proved to be accurate, but when using the .score function it was revealed that our model had a 75% accuracy on the training set and 71% accuracy on the testing set.

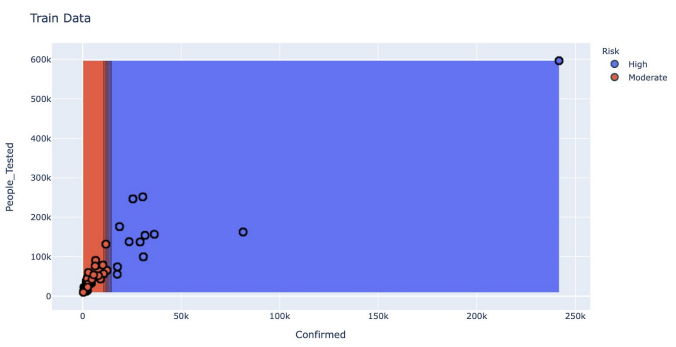
	Confirmed	People_Tested	Risk	Predicted (OVRLR) Risk
1282	30791	99727.0	High	High
2276	31652	153965.0	High	High
2477	6589	90586.0	Moderate	Moderate
415	17669	74208.0	High	High
2728	18704	176239.0	High	Moderate
533	17669	74208.0	High	High
919	1821	17676.0	Moderate	Moderate
2606	18704	176239.0	High	Moderate
220	30491	251614.0	High	High
2675	18704	176239.0	High	Moderate

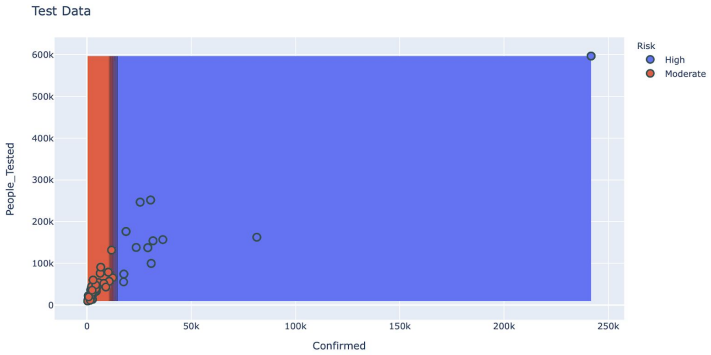
We also plotted the decision boundary for the training and test datasets for *combindedcounties.* One-vs-rest logistic regression was able to find a decision boundary between high and moderate risk states. It generalized states as high risk with high numbers of confirmed cases (>20k) and higher numbers of people tested.





Next, we attempted to use a decision tree model, but without pre or post pruning the tree we had models extremely overfit with 100% accuracy due to how specific the decision tree model was as it narrowed down to a smaller sample of events.





Finally, we examined the two different models on the original data (*combinedcounties*) to see how these models performed on the entire dataset. The logistic regression OVRLR model proved to be a better model producing a 74% accuracy score compared to the decision tree's overfit 100% accuracy.

	State	CountyName	Abbreviation	Confirmed	Deaths	People_Tested	Smoker_population	Total	Risk	Predicted (OVRLR) Risk	Predicted DT Risk
412	Georgia	Candler	GA	17669	673	74208.0	0.001999	10836.0	High	High	High
26	Alabama	Escambia	AL	4712	153	42538.0	0.000610	36748.0	Moderate	Moderate	Moderate
2917	Virginia	Tazewell	VA	8053	258	51931.0	0.000476	40855.0	Moderate	High	Moderate
1942	North Carolina	Haywood	NC	6328	187	76211.0	0.000293	61971.0	Moderate	Moderate	Moderate
148	Arkansas	Jefferson	AR	1744	38	24141.0	0.000313	68114.0	Moderate	Moderate	Moderate

Limitations

Some limitation of the analysis we did is that all the COVID-19 datasets are continuously evolving, causing data to become inaccurate even the day following the analysis. The data was collected on April 18, when the entire nation was on a stay at home order, but as of May 1, many states began reopening, causing many more people to be exposed to the virus. The number of cases and deaths will change daily especially as more states reopen nonessential businesses. With this virus, many people are asymptotic and do not get tested which also limits the accuracy of

confirmed cases. We had to assume that confirmed cases would also include those that had been exposed to the virus and were asymptotic and give us a full picture of high-risk states. Additionally, had we had access to the amount of confirmed, deaths, people tested, ect. for each county we could have created a more accurate model by using more than 50 different values since we had to base the confirmed, deaths, and people tested for each county on the totals from the entire state. Another important aspect would be to look into the demographic or layout of the city to give a better indication if impacted cities were more affected and if high-risk states that were more spread out have some external factor contributing to their high number of cases. This would've strengthened the analysis performed and provided a more holistic approach to predicting future deaths.

Ethics

The ethical dilemma present in this data was that we had to exclude counties and only looking at the state's data in aggregate. It raised the concern that by choosing to not examine how COVID-19 is affecting different regions in the state we lost valuable data that contributes to future deaths. Some concerns, for example, is how this virus is disproportionately affecting lower-income areas due to lifestyle and occupation–many lower-earning jobs have been deemed essential, such as grocery store workers and delivery drivers. In order to address the problem, it would be important to try and gather daily reports by county as opposed to gathering by the state. With this data made accessible, it would be possible to publish and combat the issue of those of lower socioeconomic statuses, usually made up of minority groups, who have been dealing with COVID-19 at a disproportionate rate and provide them with the extra resource or regulations necessary to promote health and protect their safety.

Conclusion

After completing both aspects of the analysis it is clear that predicting whether a state is high or low risk is the better question to be addressed in regards to the limitations of the evolving data. As more data becomes available this model would be applicable when adding new features to determine whether a state is moderate or high risk. More specifically new data on social distancing and the rate of a population wearing masks could be useful to track metrics that would assist in determining the risk factor. A future area of study would be predicting the risk level of the county because currently there is data on the state, but none for tracking the metrics of the virus in a specific county. This would provide valuable insight that could be applied to a more holistic study, involving new external factors that are arising each day as waves of COVID-19 progress not just throughout the entire United States, but globally.