**Predicting Weekly Infection Hotspots through Predictive Modeling**

Victoria Capobianco

Joshua Rotuna

Caroline Otten

BSAN 6070

May 3,, 2021

1. **Introduction**

***Predictive Question & Research Justification***

Throughout the past year, the COVID-19 pandemic has forced the world into lockdowns and cases have spread across different counties through the state and nation. Now, as efforts to re-open counties are enacted, how can county governments be sure it is safe to open and determine if there is a threat of new COVID-19 cases? This idea formed the basis for our research and we determined the strongest threat to a county would be whether or not that county was an infection hotspot. Our two predictive questions we used to further formulate our research are:

*Can we predict if a county will be deemed an infection hotspot?*

*What features play a role in determining if a county is a hotspot or not?*

In our research, we discovered several sources (such as Mayo Clinic, The New York Times, and the CDC) host COVID-19 tracking maps on their website which determine hotspot eligibility on a colored tier-based scale where tier is based on the average number of cases per 100,00 people. For example, California’s tier system ranges from purple to yellow, with purple being the widespread cases and yellow being minimal per 100,000 people. The issue is that this system does not take into account other factors that are affected by the virus such as the hospital capacity. Additionally, these tracking services are reported after the fact not beforehand. Therefore, our goal is to go beyond tracking the virus by analyzing a holistic picture of infection and areas it affects, such as hospitalizations, population size, and demographics, through predictive modeling. Because this is a predictive model we had to determine a timeline for how far in advance predictions could be made. We felt the daily level would be too soon and not give counties enough time to prepare for an influx in cases, while monthly prediction would be too far in advance; therefore, we settled for a weekly metric.

***Literature Review***

Before preparing our models we conducted a literature review to see if any similar studies had been conducted. We found two studies that centered around predicting infection hotspots published in 2020. The first, *An Efficient COVID-19 Prediction Model Validated with the Cases of China, Italy and Spain: Total or Partial Lockdowns?*, from the Journal of Medicine, published in May 2020, utilized the Verhulst equation in Matlab to measure infections as a function of time as the output and Initial infected population, total # of infected people, and growth rate as the inputs. Additionally, this study only utilizes data from China, Italy, and Spain. The second study was titled *Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study,* published in December 2020 for the Lancet Public Health. In this study, researchers collected self-reported data from the COVID Symptom Study App. App users were asked to self-report if they felt healthy or if they had any symptoms for 9 days. If they experienced symptoms, they had to take a PCR test, the results of which were also recorded in the app. Researchers then conducted a logistic regression using this symptom and test data to estimate rapid increases in cases across England. Based on these two studies it is clear there are already conversations within the medical community surrounding hotspot predictions. Our project adds to this conversation because we use more than reported cases and symptom-data to create a full-picture of the infection risk by including demographic information. Additionally, both of these studies used international data whereas our research is based in the US.

***Domain Understanding***

The economic and business impact of COVID-19 and rapid hotspots are astronomical. Organizations that have a stake in this issue consist of government organizations, who need to keep their constituents informed and safe regarding viral spread, hospitals and healthcare workers, who need to be aware of a potential increase in cases so they can ensure they have enough staff and supplies to aid sick patients, non-essential and small businesses, who need to plan whether or not their business will be open and operable for the upcoming future and if they will be able to generate revenue, and finally schools, who need to make plans regarding learning environments. For example, if it is predicted that a county will be an upcoming hotspot, the school would need to prepare for remote learning. At the economic level, we came across differing impacts. Economist Ryan Bourne reported the difficulty in determining a cost-benefit analysis of the COVID-19 lockdowns due to the uncertainty and ambiguity in some factors, such as the statistical value of a life, or what long-term effects would be incurred, and the difficulty in assessing these subjective values. On the other hand, two economists from Harvard reported COVID-19 will cost the US approximately $16 trillion if it ends by Fall 2021. This number was a conservative estimate based on factors such as the cost of lost economic output, premature deaths, long-term disability for individuals who recovered from the virus but still face health issues as a result, and mental health impacts. Overall, we see there are a variety of businesses and organizations that are impacted by COVID-19 and the presence of a hotspot. Our predictive model could help mediate these potential costs by aiding stakeholders in planning and removing uncertainty.

1. **Data Source & Model Planning**

In this project, the majority of the data transformation and data transformation was completed in Excel an. The original dataset required the addition of certain county features such as population stats, age, and demographics, so we mainly used VLOOKUPS in order to add the necessary columns. Next, we transformed the data from daily to weekly because this would allow more time for action regarding any prediction or decision made from the models. After doing this there was roughly 100 rows in the beginning of the time period that had missing data. We dropped these rows because they mainly came before March 2020 anyway, which was before Covid data was even being recorded.

**Feature Selection**

Numerous features were added to this dataset to observe the impact they had in causing a county to become a hotspot or not. First, we wanted to see how population and population density affected whether or not an area would or could become a hotspot. This is because we know that covid spreads easier when people are in close contact and we wanted to see if areas with people in close proximities would affect the county. Next, we wanted to see how age plays a role because we know that the elderly are more at risk. Next, we added in per capita income to see if the wealth of an area affects whether it is more likely to be a hotspot. After this, we added demographics data to our dataset because we wanted to see how different demographics might play a role in whether or not a county is considered a hotspot. We found the percent of white population, black population, native american population, asian population, Hawaiian Pacific islander population as well as other racial groups. There’s a lot of research being done currently to see whether any ethnic or racial groups are more vulnerable to covid, so we wanted to see if the percentage in each county had an effect.

Finally, a categorical column was created that labeled each county as a “hotspot” or “No Threat”. According to the Washington post who released a medical study regarding Covid hospitalizations, “Roughly 20 percent of symptomatic covid-19 patients require hospitalization” (Bernstein, 2020). Therefore, this column was calculated by multiplying the number of positive cases (symptomatic patients) by 20%. The figure that was calculated was then subtracted from the number of total beds available in each county. If there were less than 25% of beds available, the county would be labeled as a “Hotspot” for that particular time.

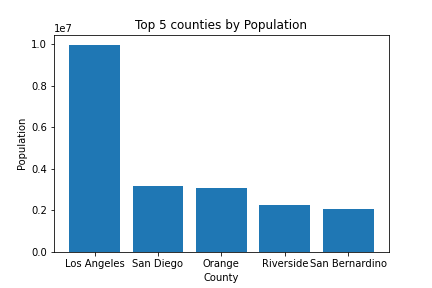
**Descriptive Statistics**

The dataset used exhibits weekly data regarding Covid statistics, such as reported cases, percent of hospital beds available and it also contains population demographics. The data is from March 29, 2020 through April 7, 2021. This dataset contains 2700 rows that each represent a specific county at different times. Each county has 18 feature columns associated with it. There are 56 unique counties, so each row provides a different scenario of Covid statistics within each county at a different time. This chart provides the average demographics, age and population stats across the entire dataset

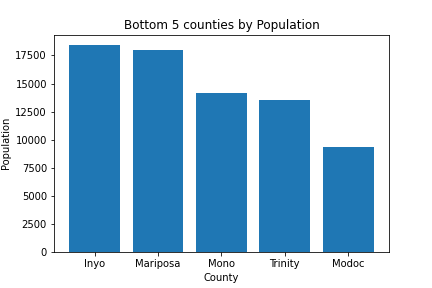
The 56 counties in our dataset observed populations that ranged from almost 10 million to just under 10 thousand. Los Angeles is the most populated county followed by San Diego and then Orange County. Modoc, Trinity and Mono County are the three least populated counties. This dataset contains counties with population densities ranging from 1.8 people per square mile to 3,575 people per square mile. The graph on the right shows the top 5 counties in terms of population density. It is interesting to note that San Francisco is the densest county but is not one of the most populated counties in California. The average age across all counties is 39.1.

Below are visualizations that further depict population statistics and demographics.

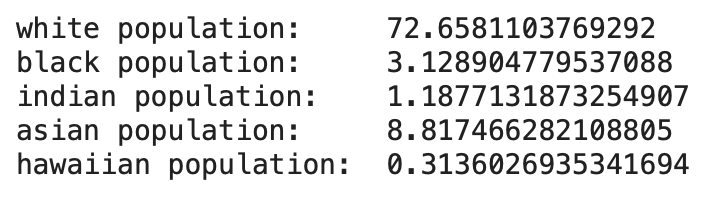
*Top 5 Populated counties:*



*Top 5 Least Populated Counties:*

****

*Average Demographics (Across entire dataset)*

******

**Dataset Assumptions**

It is important to note the assumptions associated with the entire dataset and with each model. First, we assumed that the population numbers would remain the same in each country throughout the duration of analysis. We also assumed 20% of symptomatic patients would require a hospital bed. Furthermore, we assumed that if there were less than 25% of beds available at the time, the county would be labeled a "Hotspot", if there were ample beds available, the county would be labeled as 'No Threat'.

**Performance Criteria**

For whichever model that was performed, every value, such as Accuracy, Recall, precision, and Support. However, an emphasis was placed on Recall in our model selection process because Recall is the best metric for evaluating missed positive predictions. The recall score is the number of true positives over the number of total positives it predicted. In other words, this is a metric that determines how well the model "catches" true positives. In each model, we used the GridSearch optimizer to select the parameters that would result in the best performance metrics.

Before heading into the possible models, it is important to address a class imbalance issue with our dependent variable. There were substantially more counties that were labeled as “No Threat” in that nearly 75% of the counties were labeled this way. This could cause the decision tree to create splits in which it separates out the majority class into pure groups much too early. If this happened, the model would never be able to predict an occurrence of the minority class, in this case a Hotspot. Therefore, a random sample was taken from the majority class to create an even number of cases within each class (Same number of “Hotspots and “No Threats”. The dataset created with even classes was used in each of the following models.

**Model Possibilities**

We chose three models that we figured could be used best to determine hotspots or to predict counties that could be at risk. The models selected to perform were a KNN-Nearest Neighbor Model, Logistic Regression, and a Decision Tree. We conducted a KNN model in order to predict counties similar to each other. This would allow us to insert the statistics of a certain county and find the most similar counties. If the majority of similar counties are considered HotSpots, this county either is as well, or is at a large risk of turning into a hotspot. Next, a Logistic regression was conducted to predict Hotspots. In this regression, the most important features were: Average Age, Population density, and Per capita Income. Finally, a Decision tree was conducted. This would allow us to plug in the statistics of a certain county and predict whether or not that County should be labeled as a Hotspot .

***Logistic Regression Model***

The first model we chose was a logistic regression. We deemed this appropriate for our data because our target class is binary, Hotspot vs. No Threat that we transformed into 1 and 0, while our independent variables are continuous. In conducting our model, we began by using the newton-cg solver, however, despite increasing the number of iterations, our model could not reach the lowest convergence. Therefore, we switched to the liblinear solver to get a clearer output. The final outputs were an accuracy score of 0.52, precision at 0.63, recall at 0.13, and F1 score at 0.22.

***KNN Model***

The knn Model can be used for classification. It can make predictions when there are only 2 classification outcomes, and it can be used with continuous independent variables. Therefore, it is appropriate for this data. It will calculate the nearest neighbors to a given data point, and in the hyperparameters, we can specify the number of neighbors we want to include. Each of these neighbors belong to a class and the given data point will join whichever class the majority of its neighbors are a part of. We used grid search to find the best hyperparameters for the model. We also had used random search to test different hyperparameters but only one hyperparameter changed and the accuracy, precision, recall and f1 scores stayed the same. The accuracy score was 0.6, the precision score was 0.62, the recall score was 0.62 and the f1 score was 0.62 as well. This is not as good as we would hope for and other models had higher scores all around, which is why we picked a different one as our final model.

***Decision Tree Model***

A decision tree is a very flexible algorithm and it works well with this data because it can use continuous and categorical data. In this model the inputs were continuous (Age, population, density, demographic numbers..etc). In this model, for each continuous variable the tree would determine an appropriate threshold to split at. In this specific model, the gini method was found to be the most efficient splitting criteria for this dataset. In this criteria, the model would determine its split based on whichever split results in less impurity.

After evaluating the performance metrics of each model that was conducted, the decision tree consistently outperformed the other models. The decision tree's accuracy score was roughly 67%, which is not outstanding but it is better than a random guess. After alleviating the class imbalance issue, 50% of the training data were hotspots and the other 50% were considered not to be a threat. Without the decision tree model, it would be a 50/50 guess However, the lift of this model is 17% in that it predicts Hotspots with 17% more accuracy than a guess would.

1. **Best Model & Results**

***Final Results***

For our decision tree model that we chose as our final model, our accuracy was 0.67. This is a decent score for our model. Accuracy is the # of correct decisions/total # of decisions. We know that this is too simplistic to base our decision on, as it may be misleading if you have class imbalance. However, as mentioned earlier, we took care of our class imbalance problem and have equal numbers of each class. Therefore, our accuracy score could be a good evaluation metric for our model. However, due to this measurement’s simplicity, we wanted to look at all the scores to get a better picture of how well our model performed.

Recall is the number of true positive/total number of actual positives and our recall score was 0.72. This is the score we placed the highest emphasis on, and it ended up being the highest score.

Precision is the number of True positive/total number of predicted positives and our precision score was 0.68. Both of these scores are higher than our accuracy score. We know that we need to balance the two and to look at how well we’ve done that, we can look at the F1 score. The F1 score is (2\*Precision\*Recall)/(Precision + Recall). Our F1 score was 0.7. This means we balanced our precision and recall fairly well. Looking at all of the scores, we can see that our model can make predictions with decent accuracy and with 17% more accuracy than a random guess would.

***Feature Importance***

Next we can look at the feature importance results. Average median age had the highest feature importance at 0.5, the rating was much higher than any other feature. This may be because older populations are at a higher risk, so it made sense to us that this would have the highest importance. Most of the percent of race features had very low feature importance, except for the average white percent of the population, which had the second highest feature importance. This may be due to the fact that this group had the highest percentage out of all the racial groups for the majority of the counties.

Per capita income and population density had about the same level of importance just above 0.1. Higher population density areas may be more vulnerable to covid, due to more people in the area. Although there are many covid restrictions, it can still be difficult in these highly populated areas as people still need to go to public places like grocery stores. In terms of the per capita income feature, we are unsure if lower or higher income areas were more at risk for our model. However, we found a study showing that the areas with the highest and lowest levels of poverty had reported the highest number of cases, as opposed to areas with mild levels of poverty. The authors of the study recognized that this is strange. They explained that it may be because richer areas with a high population density are infected first. Then people from neighboring, lower income areas provide work in the areas with higher wealth and are affected when these richer areas become infected. Therefore, covid spreads from the wealthier, highly populated areas to the lower income, neighboring areas.

***Limitations and Future Opportunities***

We had missing data on certain counties. This was mainly because we were doing a vlookup for our demographics columns and the data that we were combing didn’t have all the counties that were included in our original data set. We ended up having to drop these rows, but luckily we still had plenty of data to move forward. We also had missing data on the number of hospital beds available for certain rows. They were all rows that were towards the beginning of Covid back in March of 2020, so we thought that hospitals might not have been collecting that data at that point yet. Another limitation is that it was difficult to define exactly what a hotspot is for Covid. There was no exact definition and it took us some time to find a means for calculating what a hotspot is based on real life statistics.

For future opportunities, one idea we had in the beginning was to predict future hotspot for covid or influenza by combining influenza data with our covid data. This didn’t work out due to difficulties in finding data that we could combine with our covid data. However, with more time or a dataset that’s easier to combine with our covid data, we may be able to look into this question. Another opportunity would be to include additional features, such as percent of population with health insurance or percent of population above 65. This is something that we could have achieved with more time. Lastly, an idea we had was to split the label column into 3 categories, low, mild, and hotspot. This was our original idea that we had been working with for a while. However, when we made our label column, we found a huge class imbalance issue, as there was a huge amount of rows labeled low, but a lot less labeled mild or hotspot. So to fix this, we ended up combining the mild and hotspot labeled rows into one category labeled hotspot. Even after doing this we still had class imbalance so we took our random sample. In the future, more time and data, we could overcome the limitations and apply these ideas to create a more complex model.

***Conclusion***

In conclusion, we found that the decision tree model was able to outperform both the KNN and logistic regression model in terms of accuracy, precision, recall, and f1 score. We placed the highest emphasis on the recall score, as it is best at evaluating missed positive predictions. Looking at the feature importance for the decision tree model, we can see that the variable median age had the highest feature importance, which may be due to the fact that the elderly are more vulnerable to catching covid and being hospitalized due to covid. The race features did not have a high feature importance, except for the white percent of population variable. The population density and per capita income had the third and fourth highest feature importance and were around the same level of importance. With a lift of 17%, our model can make predictions with 17% higher accuracy than a 50/50 guess. This model can help stakeholders, such as government organizations and schools, to better plan covid procedures in areas predicted to become a hotspot.

Sources:

<https://news.harvard.edu/gazette/story/2020/11/what-might-covid-cost-the-u-s-experts-eye-16-trillion/#:~:text=A%20pair%20of%20Harvard%20economists,and%20contact%20tracing%20that%20would>

<https://www.cato.org/blog/cost-benefit-analysis-lockdown-very-difficult-do-well>

<https://www.mdpi.com/2077-0383/9/5/1547>

<https://doi.org/10.1016/S2468-2667(20)30269-3>

<https://data.census.gov/cedsci/table?q=United%20States&t=-00%20-%20All%20available%20races%3AAge%20and%20Sex%3AHealth%20Insurance%3ARace%20and%20Ethnicity&g=0400000US06,06.050000&y=2019&tid=ACSSPP1Y2019.S0201&hidePreview=true>

[https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state/resource/1be1e43c-b4b2-4002-afb6-340bbcc85bb](https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state/resource/1be1e43c-b4b2-4002-afb6-340bbcc85bbf)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7756168/