Kristin Levine

Special Project

August 2020

Race/Ethnicity -- Reopening Dates and Masking Requirements

For my special project I wanted to look at reopening dates and masking requirements on different racial and ethnic groups. I selected four features: state, number of cases on June 15, reopening date, and date of masking order. The target feature was whether or not the percent increase in cases from June 15 to July 20 was >= 100%.

**Cleaning the Data**

June 15 was the first day the health equity team started collecting race/ethnicity data, so I decided to use that as my starting point. I decided to focus on six groups: White, Black, Asian, Hispanic, American Indian/Alaska Native (AIAN), and Native Hawaiian/Pacific Islander (NHPI.)

Since reopening dates and masking requirements were reported at the state level, I started by splitting the county and state into separate columns. I then cleaned the data, removing commas, symbols, etc. I summed the totals for each race/ethnic group for each state, then combined those numbers into one table I called ‘june\_total.csv’. Note: we only had data for 32 states; of those, we had no starting data for three of those states. So that left me with June race/ethnic data for 29 states.

I did the same thing for the July numbers. By July 20, we were collecting data for 35 states, but we had no ending data for 5 of those states.

I collected the state reopening data and the mask order info from the New York Times and CNN: <https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html?register=email&auth=register-email> and <https://www.cnn.com/2020/06/19/us/states-face-mask-coronavirus-trnd/ind>

and combined them into a table called: ‘Reopen\_Mask.csv’.

For each of the six groups, I then created a clean file called, for example, ‘white\_data.csv’. This table included the following columns: state, numbers on June 15, numbers on July 20, reopening dates, mask mandate dates, percent change from June 15 to July 20, and whether or not the percent change in case numbers was >= 100. I eliminated any state for which we had 0 cases recorded for either June or July. This left me with 26 states. I used these clean files to run my Decision Tree and Random Forest analysis.

**June 15 to July 20, 2020**

**Decision Tree**

|  |  |  |
| --- | --- | --- |
| **Race** | **DT Accuracy** | **Most Imp Feature** |
| White | 62.5 | Numbers on June15 |
| Black | 71.4 | State |
| Asian | 71.4 | Numbers on June15 |
| Hispanic | 50.0 | Number on June15 |
| AIAN | 33.3 | Numbers on June15 |
| NHPI | 50.0 | State |

Interestingly, knowing the state, June numbers, reopening, and mask policies did NOT allow me to predict with any great accuracy using Decision Tree/Random Forest algorithms whether or not a state’s case numbers would have increased by more than 100% over the time period of this study. The one exception was for Black racial groups, where the random forest algorithm was extremely accurate.

**Random Forest**

|  |  |  |
| --- | --- | --- |
| **Race** | **RF - All Features** | **Most Imp Feature** |
|  | Accuracy |  |
| White | 62.5 | June15 |
| Black | 100 | State |
| Asian | 71.4 | June15 |
| Hispanic | 50.0 | June15 |
| AIAN | 33.3 | June15 |
| NHPI | 33.3 | State |

![A screenshot of a cell phone

Description automatically generated]()

Feature Importance for Black -- Random Forest Data

Also, the number of cases on June 15 and/or the state appeared to be MORE important features in predicting the percent increase in cases rather than the date of a reopening or masking requirement.

All of the dates of “reopening” were before June 15, some many weeks before, plenty of time for new cases to start circulating. Masking requirements, as a whole, came much later than reopening dates, some perhaps after cases had already begun to rise.

There is so much more that could be done with this data! It would be interesting to look at if the opposite were true – could you predict whether or not a state was likely to have a masking requirement by how much their numbers were increasing in the weeks before it was issued? If so, it might be that these masking requirements are coming too late to prevent widespread transmission in many areas.

We are also missing data for half the states – more complete data, collected over a longer period of time, might give us more accurate results. We only started collecting race/ethnicity data AFTER the reopening/masking requirement.

Special Project Files:

<https://github.com/kristinslevine/broadstreet_special_project>

Original CSV Files:

* June\_15.csv
* July\_20.csv
* Reopen\_Mask.csv

Python Files:

* June\_data and July\_data create clean csv files (June\_total/July\_total)
* Each file for the 6 race/ethnicity groups creates a clean csv file entitled [race]\_data
* [race]\_DT runs the Decision Tree algorithm for that race
* [race]\_RF runs the Random Forest algorithm for that race