Covid-19 Infection and Mortality Modeling

Capstone II Final Report
Springboard data science program

Zheng Zhang, Ph.D.
September 17, 2020

Introduction

Covid-19 is a viral disease caused by a novel coronavirus that emerged in Wuhan, China in December 2019. Since then, the disease has been spreading across the globe. As of August 27, 2020, it had infected at least 24 million people worldwide and more than 822,000 had died from the complications of the disease. The United States had been one of the hardest hit countries with more than 5.9 million cases and more than 180,000 deaths. In an effort to control the spread of the disease, the vast majority of the countries had imposed mobility control in the form of mandatory stay-at-home order and had shut down domestic and international travel by March 2020. In addition, large sectors of economy had been shut down, including travel, entertainment, and education. As a result of those disease mitigation matters, the world economy had been hard-hit and unemployment rates skyrocketed. As the world is desperate for the understanding of the virus and the disease it causes, disease modeling is critical to eventually control the spread of the disease.

Problem Statement

We seek to predict Covid-19 hotspots based on movement data, health ranking and social vulnerability index. Secondly, we want to predict Covid-19 deaths based on the same set of predictors.

Datasets

A. Dataset Description

The datasets were downloaded from www.kaggle.com . Specifically,

1. Daily confirmed Covid-19 cases between 1/22/2020 and 4/28/2020 in 3130 US counties with a total of 313,110 observations and 9 features.
2. Daily Covid-19 related deaths between 1/22/2020 and 4/28/2020 in 3130 US counties with a total of 313,110 observations and 9 features.

3. US county health rankings in 2020 in 3142 counties with 3193 observations and 507 features.
4. CDC social vulnerability index in 2016 in 3142 us counties with 3142 observations and 127 features.
5. Google mobility data between 02/15/2020 and 04/11/2020 in 2818 US counties with 155,060 observations and 9 features.

B. Missing Data Handling

1. There are significant amounts of missing data in the mobility dataset, especially on residential, parks and transit station mobility with more than 50% missing. We decided to drop those variables from the analysis. The missing data on retail, workplaces and grocery mobilities are between 6-17%, and we replaced those missing values with the median.
2. Missing entries in the health ranking data are moderate at less than 10% and were left as they were.
3. There were no missing data in the covid-19 case, covid-19 death, and social vulnerability datasets.

C. Data Wrangling:

1. Covid-19 cases and deaths: we deleted non-specific counties from the data and kept the number of confirmed cases and deaths from the last date (2020-04-22) for each county.
2. Mobility data: we deleted non-specific counties from the data and calculated the minimum and maximum value for the retail, grocery and workplaces mobilities for each county.

D. Construction of the Combined dataset

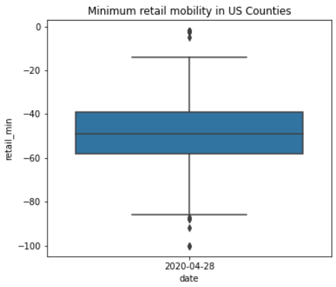
The covid cases, death, mobility, social vulnerability and health ranking datasets are merged on state and county to create the combined dataset with 2706 counties and 31 variables.

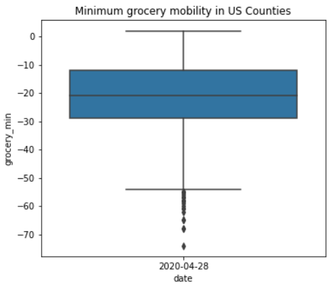
Exploratory Data Analysis

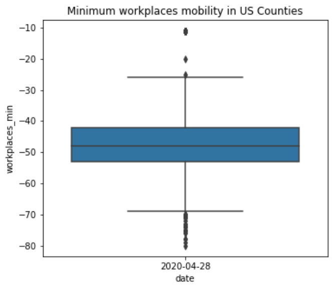
A. Distribution of covid-19 and mobility variables

1. Covid-19 cases as of 04-28-2020: Of the 2706 counties, the median covid-19 cases is 20, however, the max case is 49,929. More than half had no death, however, the max number is 5281.
2. The retail mobility had fluctuated wildly, from a complete shut-down of -100% to an increase of 226%. The patterns are the same for the other two mobility variables.

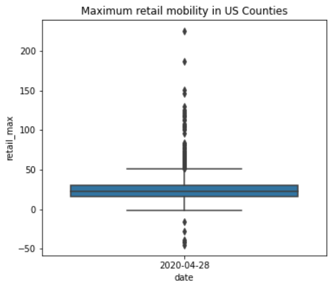
Minimum mobility:

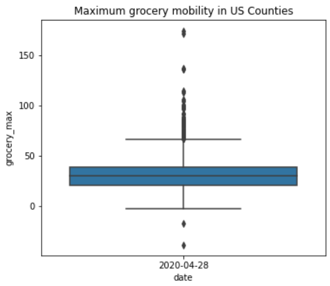

Minimum retail mobility in US Counties

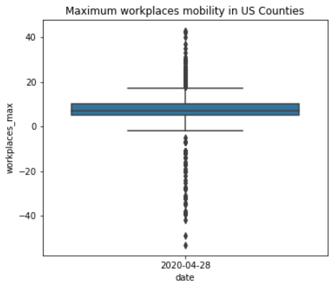

Minimum grocery mobility in US Counties


Minimum workplaces mobility in US Counties

Maximum mobility:


Maximum retail mobility in US Counties

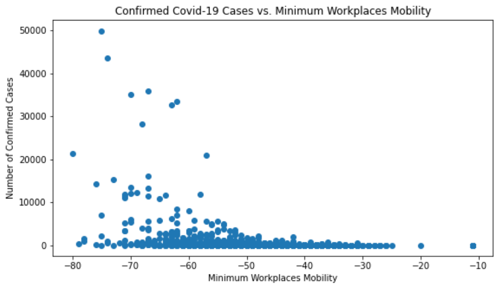

Maximum grocery mobility in US Counties

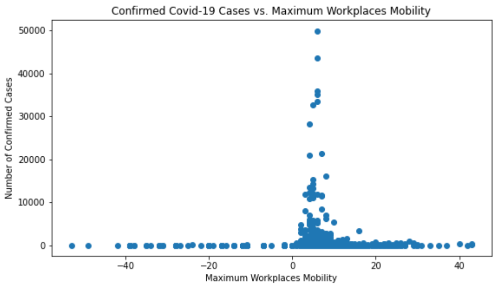

Maximum workplaces mobility in US Counties

B. Relationship between mobility and covid cases:

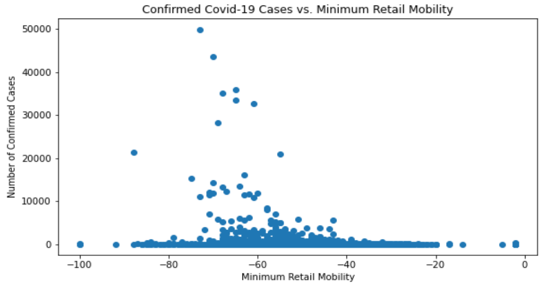
There was no obvious pattern of relationship between the mobility and active cases.

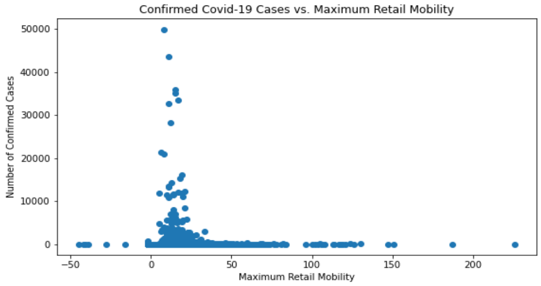
1. Workplaces mobility vs. confirmed cases


Confirmed Covid-19 Cases vs. Minimum Workplaces Mobility

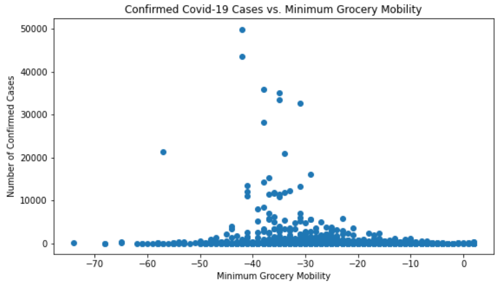

Confirmed Covid-19 Cases vs. Maximum Workplaces Mobility

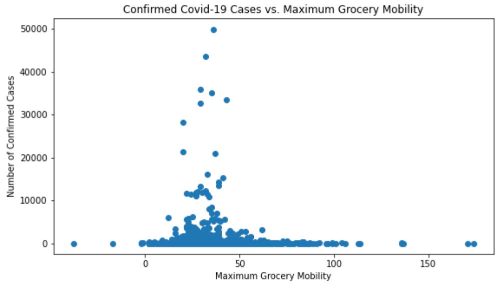
2. Retail mobility vs. confirmed cases





3. Grocery mobility vs. confirmed cases





In-Depth Data Analysis

We have decided to predict counties with high covid-19 cases, which are defined as the top 50% of the counties in regards to the covid-19 confirmed cases as of April 28, 2020. In addition, counties are also classified based on their covid-19 deaths, those with death counts in the top 50% are classified as positive.

We will use machine learning to predict the covid 19 hotspots and high death counts. We have retained 25 predictors to build the model, and they belongs to three groups:

1. Mobility(6 features): grocery min, grocery max, retail min, retail max, workplaces min, workplaces max
2. Health ranking (15 features): number of deaths per year, years of potential life lost rate, % poor health, % low birthweight, % smokers, % driving death with alcohol, % obesity, % physical inactive, % excessive drinking, % uninsured, primary care physician rate, % with severe housing problem, life expectancy, % homeowners, % rural
3. Social vulnerability index (4 features): % poverty, % unemployment, % > age 65, % minority

After deleting counties with missing data in any of those 25 features, there are 2588 counties left.

First we looked at the correlation matrix (heatmap, Figure 1) to check on potential multicollinearity issue:

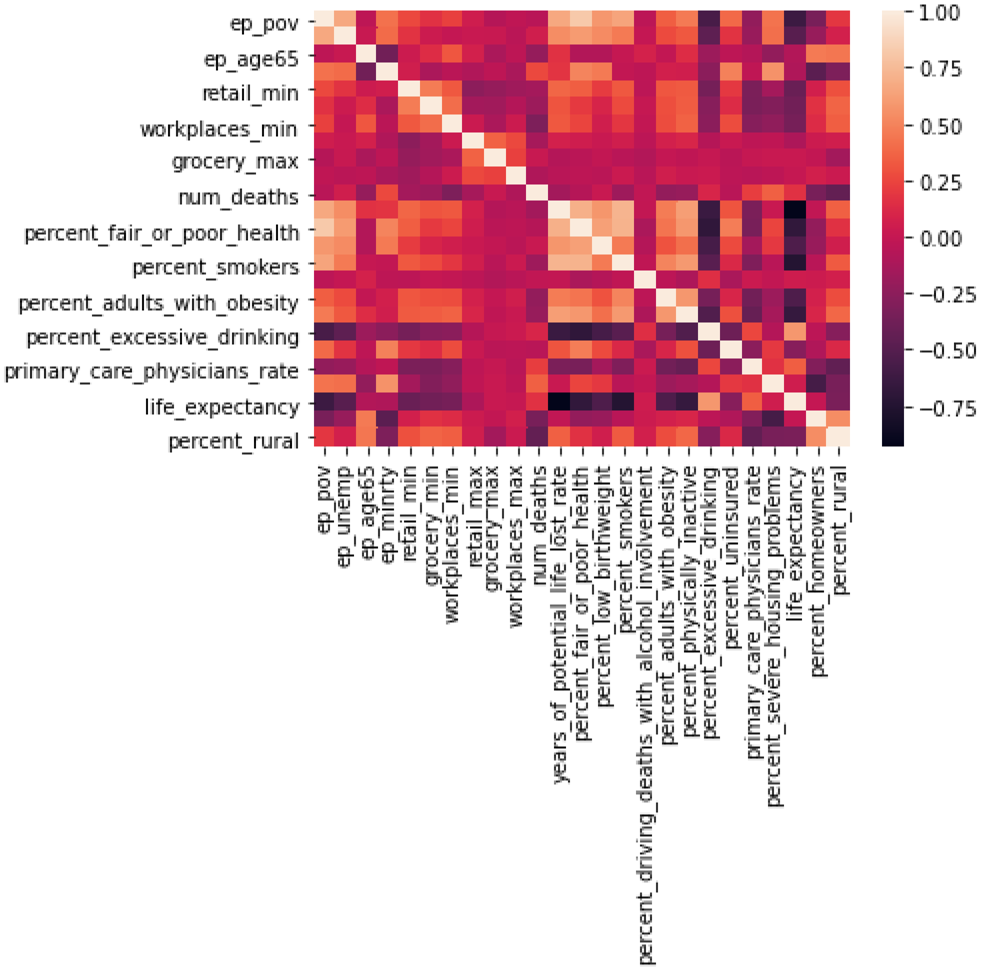


Figure 1: The heat map of the features in the prediction model

Observation: There are some degrees of correlations, but overall the multicollinearity is not a big concern.

Predict Covid Hotspots with Machine Learning Techniques

We employed several machine learning algorithms to build the prediction model. The model building process using different machine learning algorithms share some of the following common steps:

1. Split the dataset into a 70% training set and 30% test set, stratified by the case status (high vs. low).
2. Some algorithms have hyperparameters, where GridSearch has been used to identify the best choice for such hyperparameters (hyperparameter tuning) by using 5-fold cross-validation.
3. By design, there is an even distribution between the high case counties and low case counties, hence the data imbalance is not an issue here.

Results from the machine learning algorithms:

	Hyperparameter Tuning	Sensitivity	Specificity	Misclassification
Confirmed Cases				
Logistic Regression		73%	86%	20%
Random Forest	n=94	79%	90%	16%
Bagging		81%	74%	23%
AdaBoost		78%	86%	18%
Deaths				
Logistic Regression		67%	78%	28%
Random Forest	n=92	71%	77%	26%
Bagging		66%	77%	28%
AdaBoost		72%	74%	27%

As shown above, we compared the performance of four different algorithms: Logistic regression, random forest, bagging and AdaBoost.

For predicting cases, of those four, random forest achieves the highest overall accuracy at 16%, with a sensitivity at 79%, and a highest specificity at 90% (Figure 2). The AdaBoost method offers the second highest performance behind random forest.

The performance of all four algorithms to predict covid-19 deaths are less than satisfactory, with overall misclassification rates between 25-30%.

Overall, we will recommend random forest to predict covid-19 cases and deaths. The most important predictors are: number of deaths, percentage of rural, minimum workplace mobility, estimated percentage of seniors and minority (Figure 3).

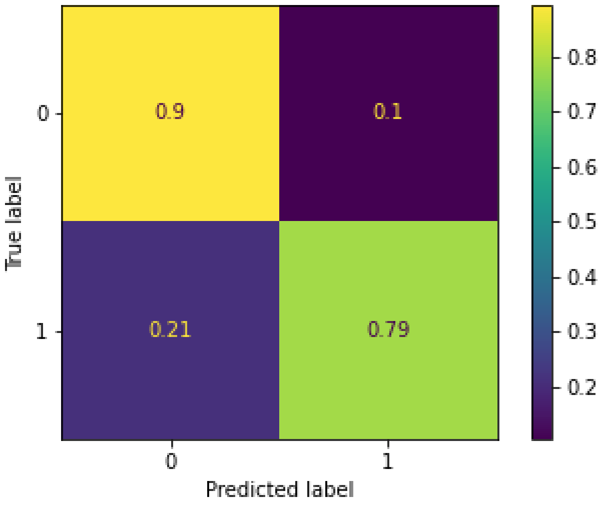


Figure 2: The confusion matrix of the random forest model to predict Covid-19 cases.

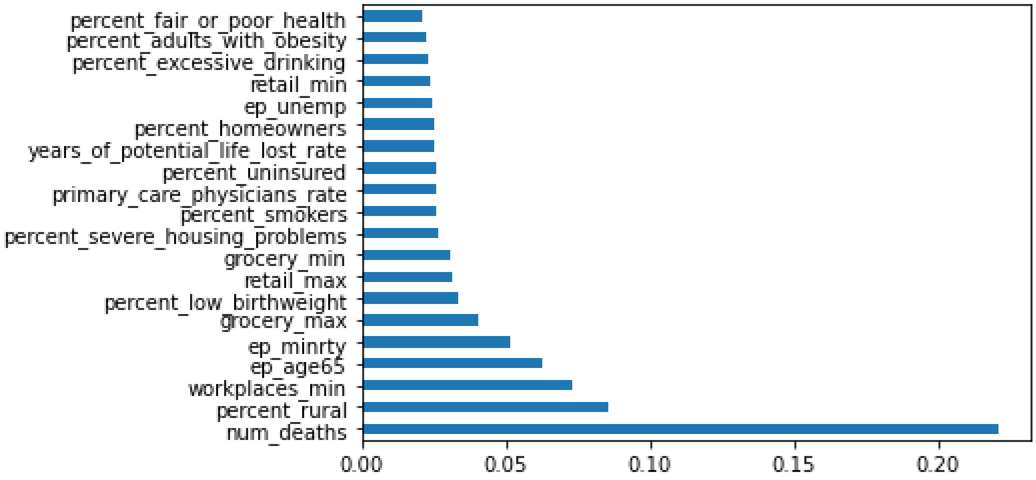


Figure 3: Feature of importance of the random forest model to predict Covid-19 cases.

Percent of the Covid-19 Cases from Correctly Predicted Counties

Although random forest has only identified 79% (308) of counties with higher covid-19 cases, it identified a total of 311,306 cases, which is 97% of all the cases in the entire dataset and 98% of all the cases in the high count counties. The median Covid cases for the high Covid counties are 137 vs. 41 for the correctly identified ones vs the missed ones (Figure 4).

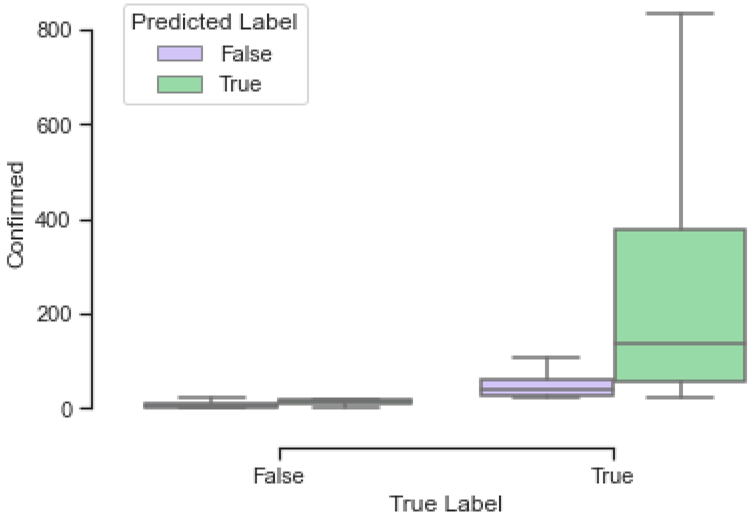


Figure 4: Predicted vs. True Label of Counties with High Covid Cases

Similarly, although random forest has only identified 71% (277/390) of counties with higher covid-19 deaths, it identified a total of 12,026 deaths, which is 98% of all the deaths in the entire dataset. The median Covid deaths for the high Covid death counties are 5 vs. 1 for the correctly identified ones vs the missed ones (Figure 5).

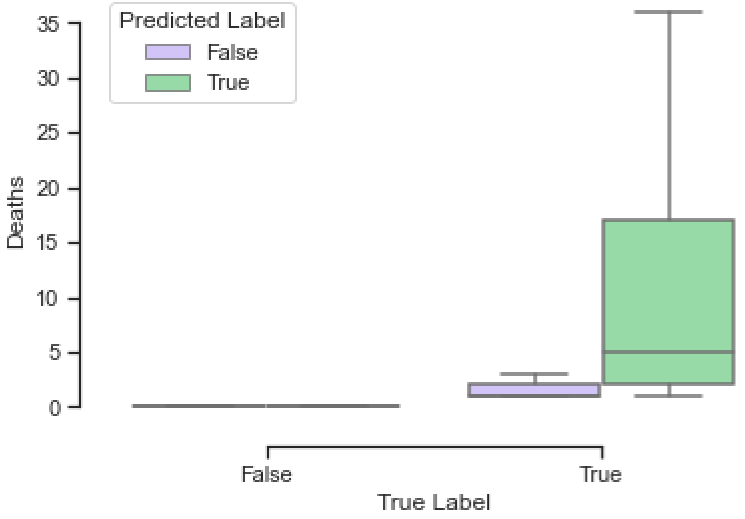


Figure 4: Predicted vs. True Label of Counties with High Covid Deaths

Conclusion

Of all the machine learning techniques that we investigated in this project, random forest was able to predict 79% of US counties with high covid-19 cases reported, which account for the vast majority(311,306, 97%) of all those cases. The median number of covid-19 cases is 137 in those correctly identified counties as of April 28, 2020. Random forest model was also able to identify over 98% of covid-19 deaths. It was to our surprise that a model based on movement data, health status and social vulnerability index was able to predict covid-19 hotspots and total covid-19 deaths. Since the top identified features of significance were number of deaths, percentage of rural, minimum workplace mobility, estimated percentage of seniors and minority, it may imply age, race, geographic area, overall health and workplace mobility are important predictors for covid-19 hotspots prediction in the early phase of the pandemic in the US.