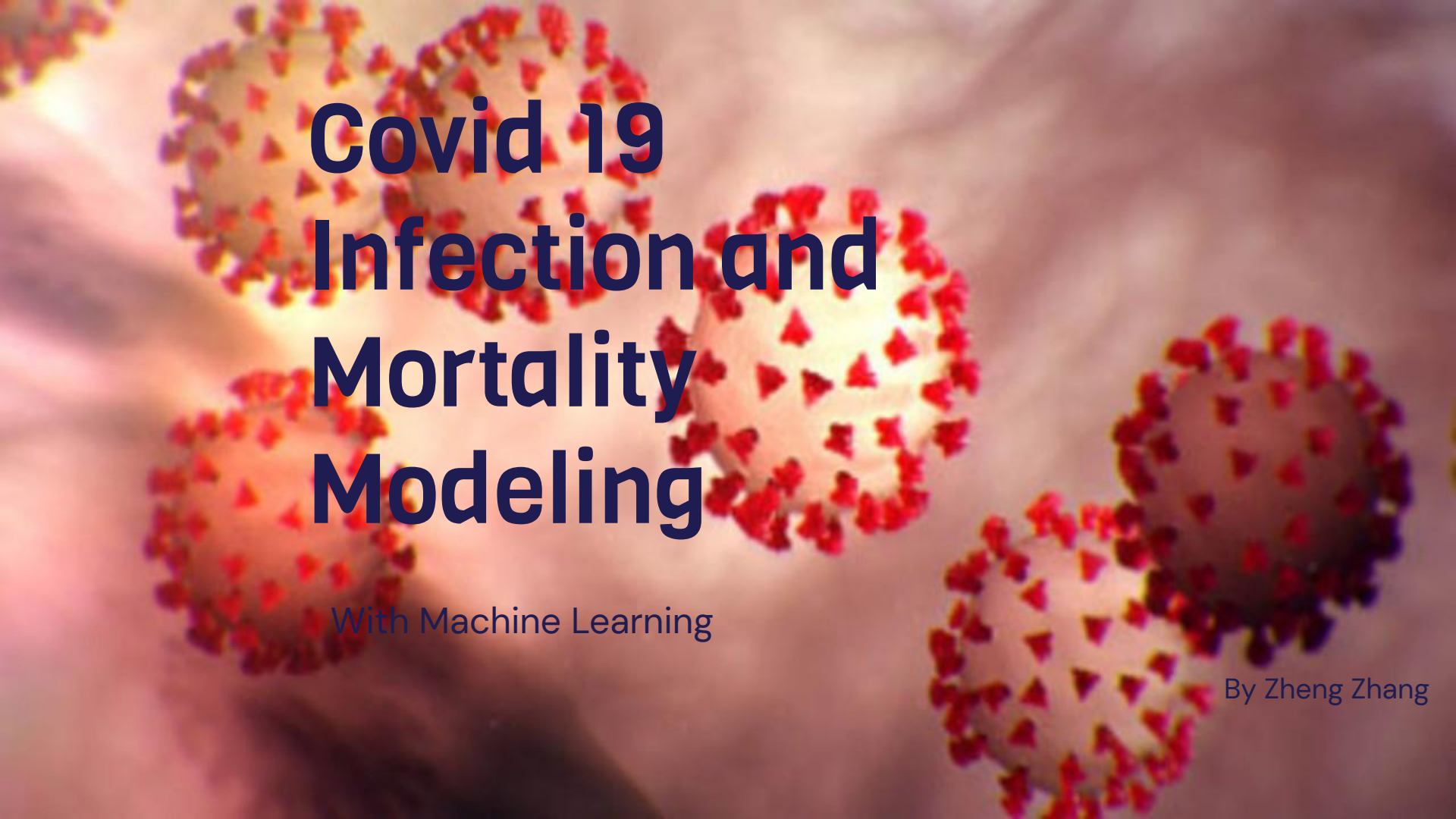


Covid 19 Infection and Mortality Modeling

With Machine Learning

By Zheng Zhang

DATASET & OVERVIEW

- Covid-19 is a viral disease caused by a novel coronavirus that emerged in Wuhan, China in December 2019. As of June 30, 2020, it had infected at least 10 million people and more than 500,000 had died from the complications of the disease. The United States had been one of the hardest hit countries with more than 2.6 million cases and more than 120,000 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.
- We seek to predict Covid-19 hotspots and mortality based on movement data, health ranking and social vulnerability index.
- We will build our prediction based on five public available datasets downloaded from www.kaggle.com.

The Value Proposition

- The ability to accurately identify the next hotspot of the infection can provide an important tool for public health decision making.

  - ❖ Behaviour modification is critical to control the spread of the disease.
  - ❖ Targeted public health intervention can reduce the burden of the disease as well as the mortality rate.

LOADING DATA

LOADING DATA

- The datasets were downloaded from Kaggle's website, which include covid-19 cases and deaths, google mobility data, US county health rankings in 2020 and CDC social vulnerability index in 2016.
- 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.
- 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.
- US county health rankings in 2020 in 3142 counties with 3193 observations and 507 features.
- CDC social vulnerability index in 2016 in 3142 us counties with 3142 observations and 127 features.
- Google mobility data between 02/15/2020 and 04/11/2020 in 2818 US counties with 155,060 observations and 9 features.

EXPLORATION

CLEANING DATA

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.
4. 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.
5. 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.

Exploratory Data Analysis

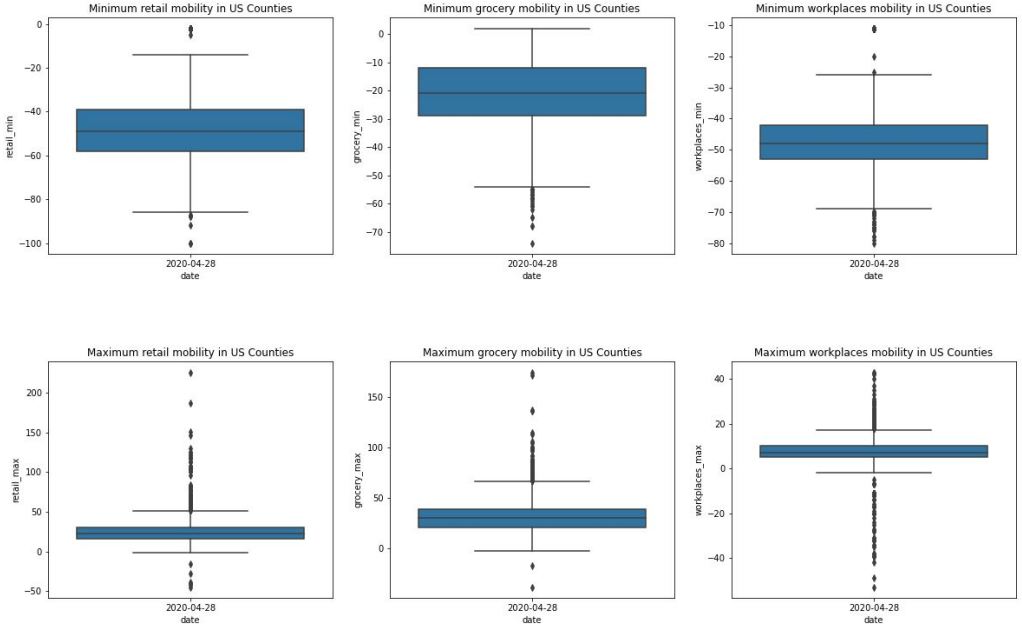
1. The covid cases, death, mobility, social vulnerability and health ranking datasets are merged on state and county to create the combined dataset with 2706 counties and 31 variables.

2. 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.

3. We looked at change in the mobility data and calculated minimum and maximum mobility for retail, grocery and workplaces.

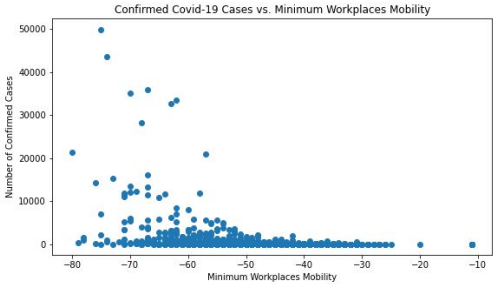
4. No obvious pattern of relationship between the mobility and active cases.

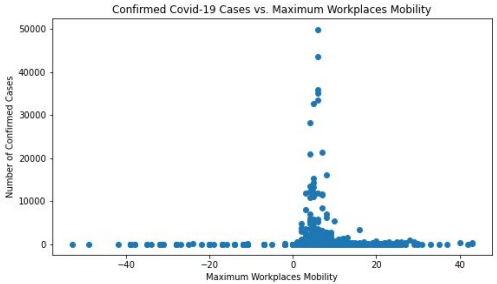
Minimum and Maximum Mobility



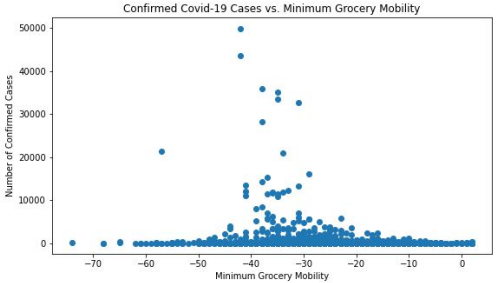
- The retail mobility had fluctuated wildly, from a complete shut-down of -100% to an increase of 226%. The patterns are the same for the other two mobility variables.

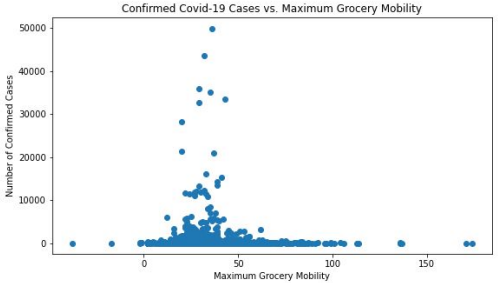
Workplaces Mobility vs. Covid-19 Cases



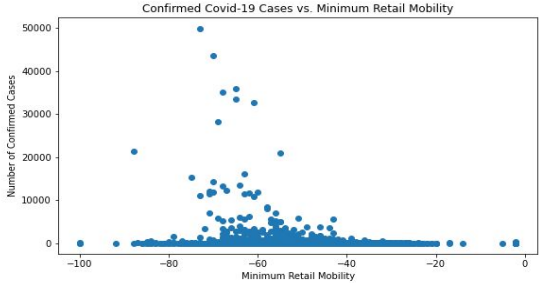


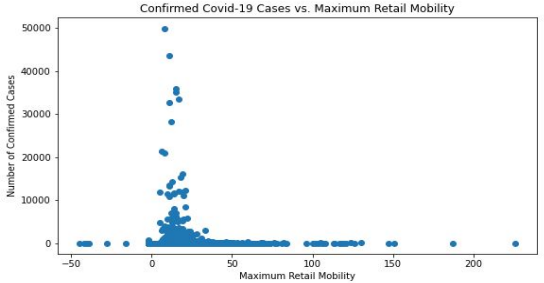
Grocery Mobility vs. Covid-19 Cases





Retail Mobility vs. Covid-19 Cases



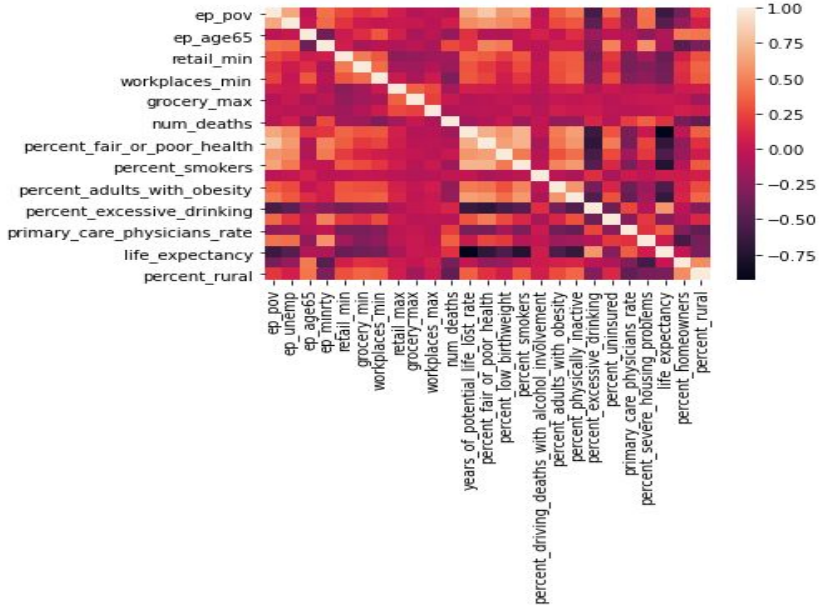


Final List of Features for Machine Learning

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

- Total 2588 counties with 25 features.

Check for Multicollinearity



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

COVID-19 Hotspot Prediction

Modelling Covid-19 Cases

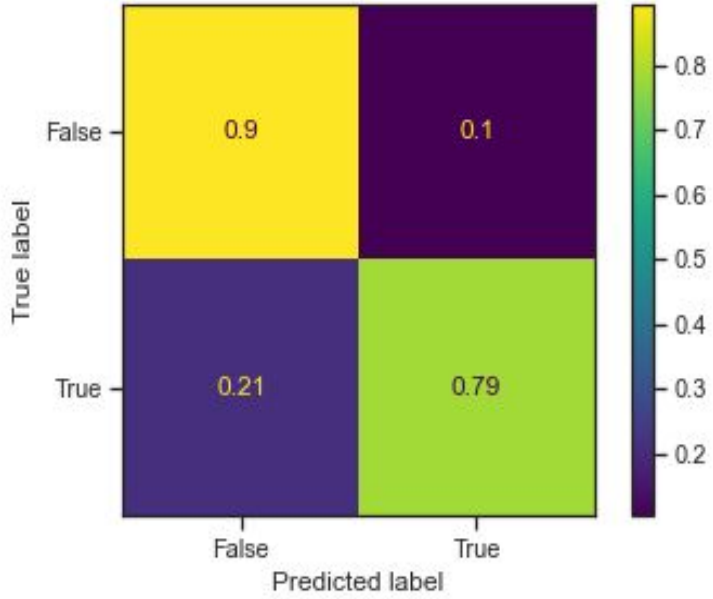
- Classify the 2588 US counties into high covid-19 counties (those with covid-19 cases at or above the median of 22), with 1304 in the high category, 1284 low.
- Split the dataset into a 70% training set and 30% test set, stratified by the covid-19 high/low cases status.
- A total of four algorithms were executed: Logistic regression, random forest, bagging and AdaBoost.
- 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.
- All models are run with the same random state when applicable

Covid-19 Hotspot Prediction

Model	Sensitivity	Specificity	Misclassification rate	Tuning parameter
Random Forest	0.79	0.90	0.16	N=94
Logistic Regression	0.73	0.86	0.20	
Bagging	0.81	0.74	0.23	
AdaBoost	0.78	0.86	0.18	

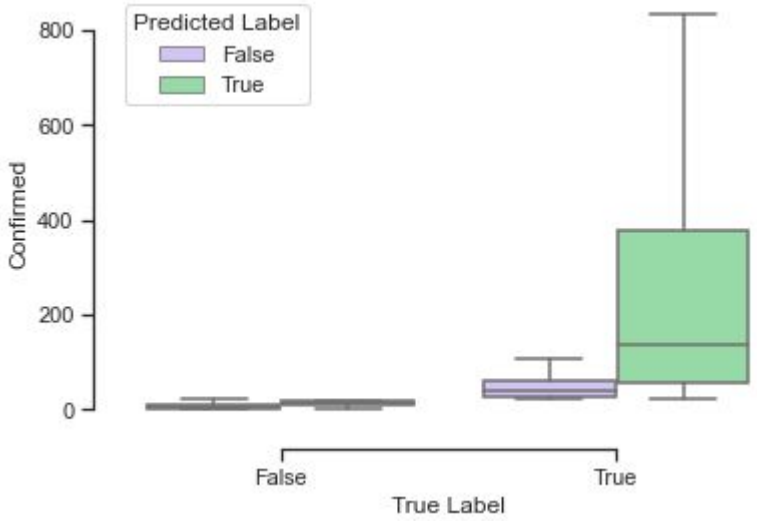
- Random forest has the best performance with AdaBoost being the second best.

CONFUSION MATRIX (Random Forest)



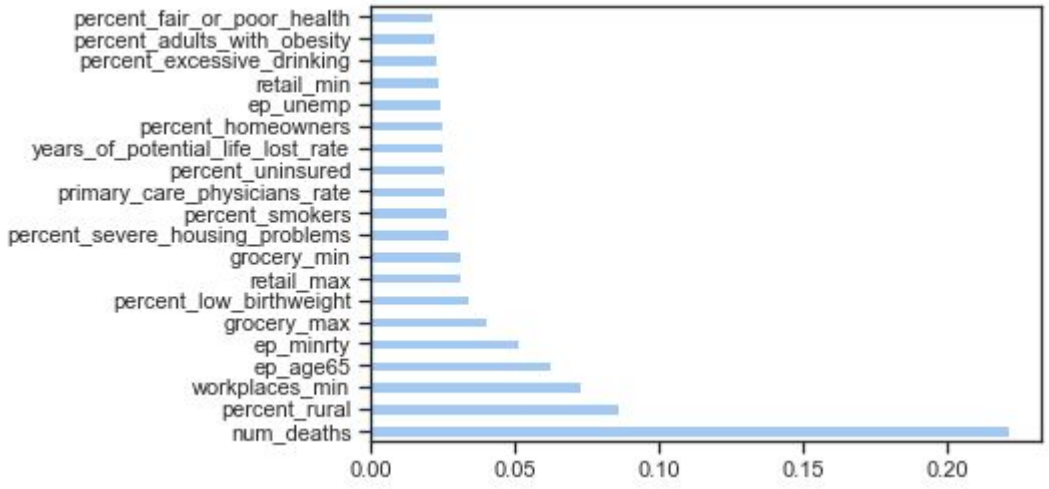
- Correctly labelling 79% of counties with high covid-19 cases.
- It identified a total of 311,306 cases, which is 97% of the cases in the entire dataset and 98% of all the cases in the high count counties.

Predicted vs. True Label: Covid-19 Cases



- Random forest has identified 79% of counties with at least 22 Covid-19 cases, with median case of 137, compared with the rest 21% missed one with a median of 41.

Feature of Importance: Covid-19 Cases



- Random forest has identified number of death, percent of rural, workplaces minimum mobility, estimated percentage of seniors and minority as the most important features.

Covid-19 Death Prediction

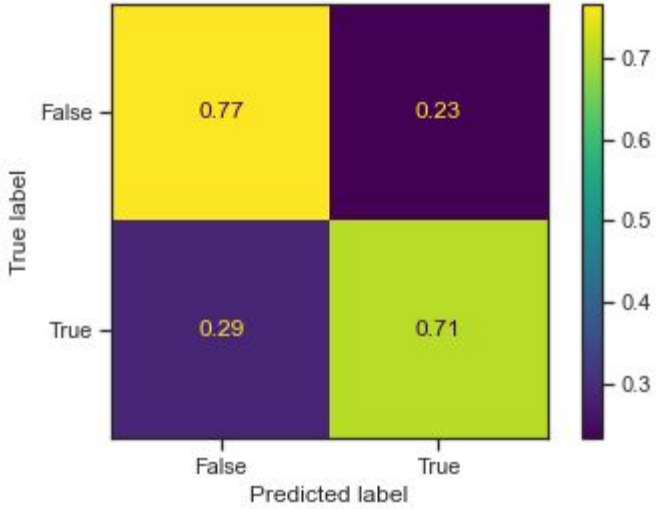
Modelling Covid-19 Deaths

- Classify the 2588 US counties into high covid-19 death counties (those with at least 1 covid-19 death), with 1299 in the high category, 1289 low.
- Split the dataset into a 70% training set and 30% test set, stratified by the covid-19 high/low cases status.
- A total of four algorithms were executed: Logistic regression, random forest, bagging and AdaBoost.
- 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.
- All models are run with the same random state when applicable

Covid-19 Death Prediction

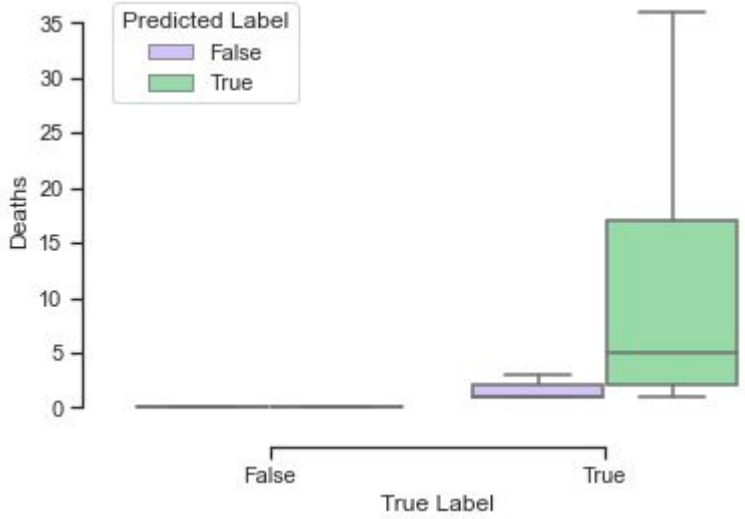
Model	Sensitivity	Specificity	Misclassification	Tuning parameter
Logistic Regression	0.67	0.78	0.28	
Random Forest	0.71	0.77	0.26	N=81
Bagging	0.66	0.77	0.28	
Adaboosting	0.72	0.74	0.27	

CONFUSION MATRIX (Random Forest)



- Correctly labelling 71% of high covid-19 deaths.
- It identified a total of 12,026 deaths, which is 98% of all the deaths in the entire dataset.

Predicted vs. True Label: Covid-19 Deaths



- Random forest has identified 71% of counties with at least one Covid-19 death, with median death of 5, compared with the rest 29% missed one with a median of 1.

CONCLUSION

SUMMARY

1. 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.
2. Random forest model was also able to identify over 98% of covid-19 deaths.
3. It was to our surprise that a relative simple model based on movement data, health status and social vulnerability index was able to predict covid-19 hotspots and total covid-19 deaths.
4. 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.