EE 510: Machine Learning Theory & Algorithms **Spring 2020**

Mini Project 2

Due: May 15, 2020, 11:59PM PT

Student Name: Andrew Huffman *Instructor Name: John Lipor*

Problem Description

In this project I worked with the COVID19 Week 5 dataset [1] from Kaggle. This dataset includes a training set and a test set, but late submissions for the competition were not allowed, which made the test set unusable. So, I used the last two weeks of data from the training set as my test set. The training set included countries/regions, counties, provinces/states, population, weights (proportional to population), dates, targets (confirmed COVID19 cases and COVID19 fatalities), and the target values corresponding to these features for January 23, 2020 through April 26, 2020 . The test consisted of the aforementioned features for April 27, 2020 through May 10, 2020. My goal was to train a model to predict the daily number of confirmed cases and fatalities for each data source (e.g. country or county). I implemented random forest regression and extra trees regression. A model that could accurately predict COVID19 cases and fatalities would obviously be useful, but especially if that model could map controllable features, such as level of social interaction or handwashing frequency, to changes in the spread of the virus and resulting fatalities.

Exploratory Data Analysis (EDA)

I found corochann's tutorial [2] helpful in getting started with exploratory data analysis. I focused on comparing the spread of the virus and associated fatalities geographically. I explored case counts, fatality counts, infection rates (as a percentage of population), and mortality rates (fatalities divided by cases) by country, by state with the U.S., and by county within Oregon.

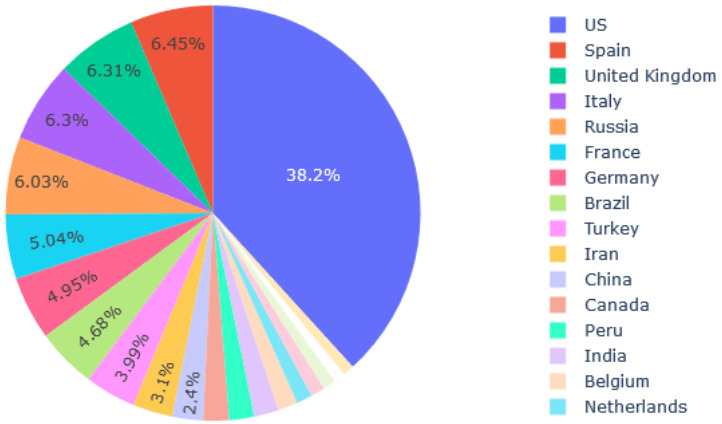

Figure 1: Top 20 Countries by Case Count

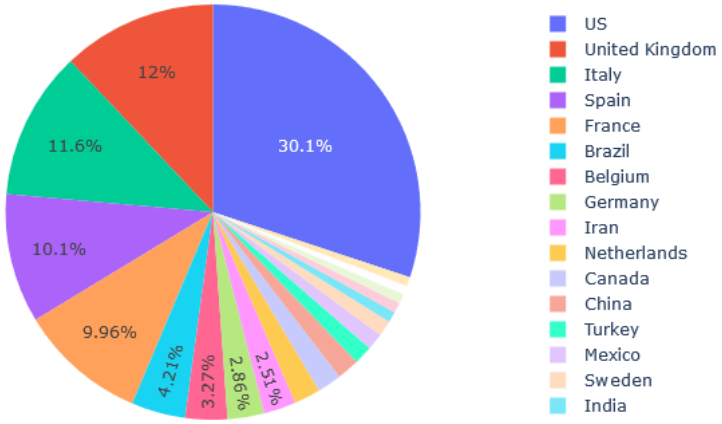

Figure 2: Top 20 Countries by Fatality Count

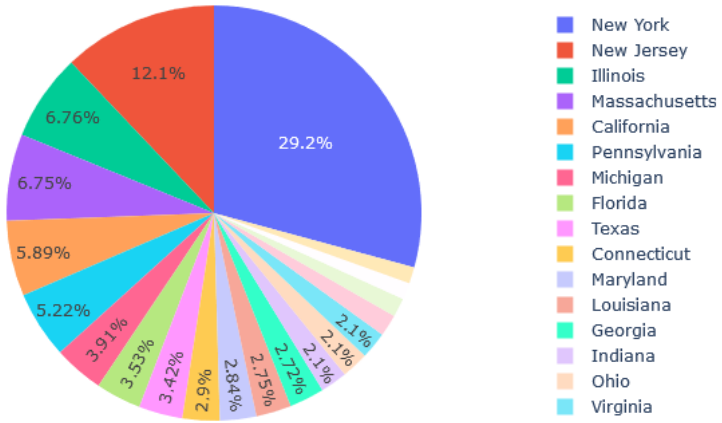

Figure 3: Top 20 States by Case Count

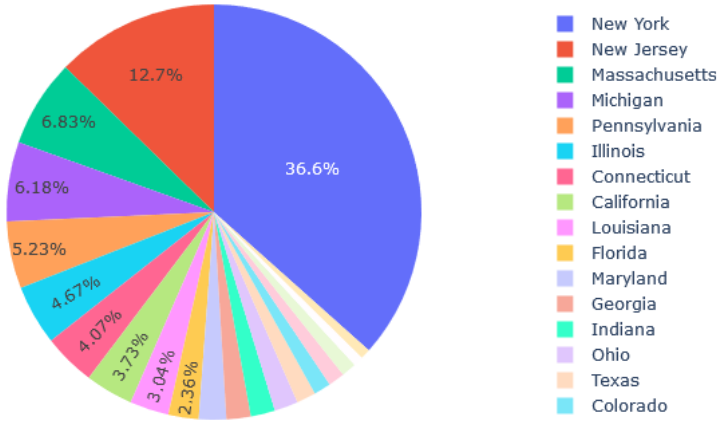

Figure 4: Top 20 States by Fatality Count

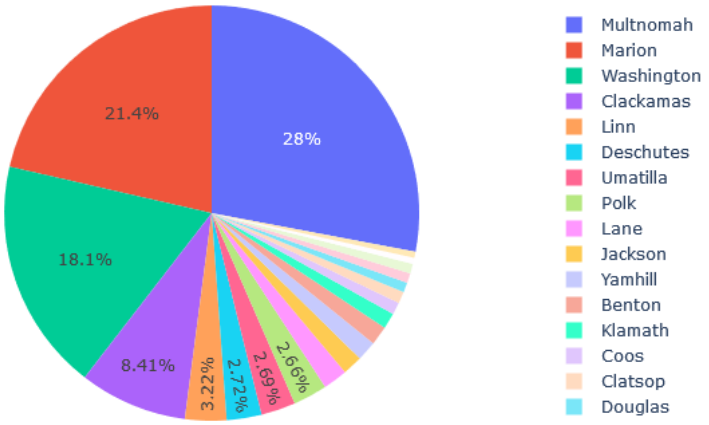

Figure 5: Oregon Counties by Case Count

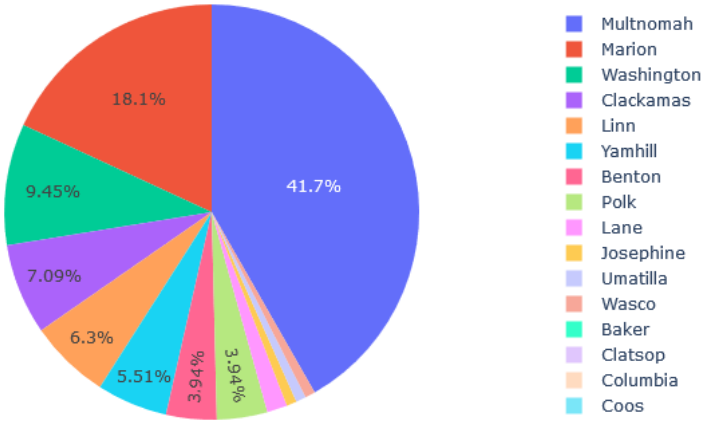

Figure 6: Oregon Counties by Fatality Count

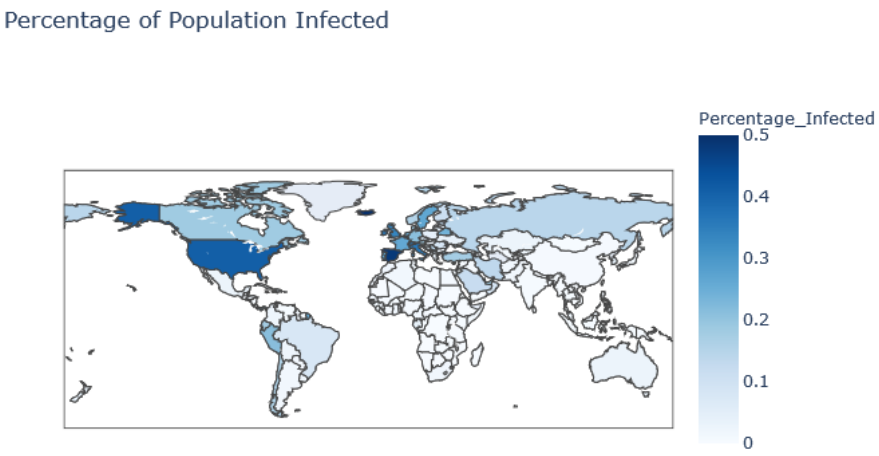

Figure 7: Global Infection Rates

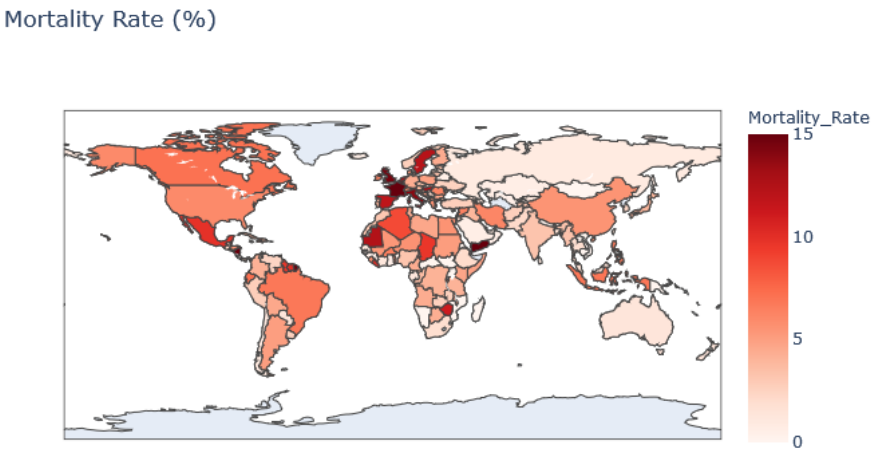

Figure 8: Global Mortality Rates

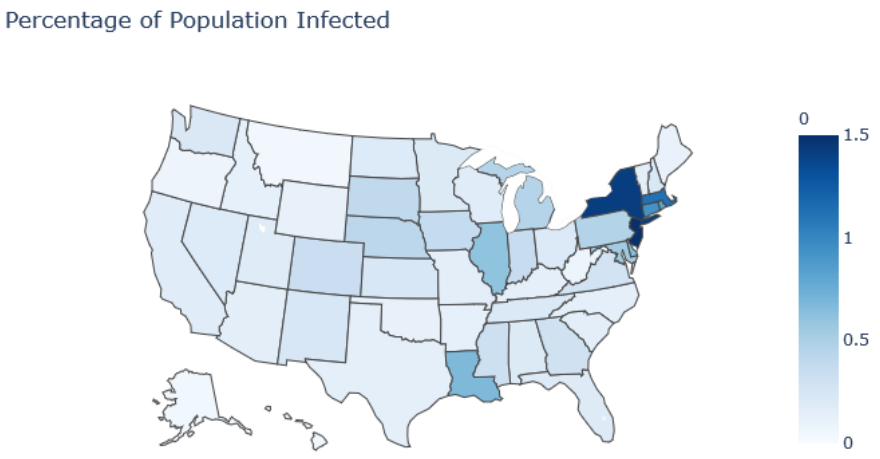

Figure 9: State Infection Rates

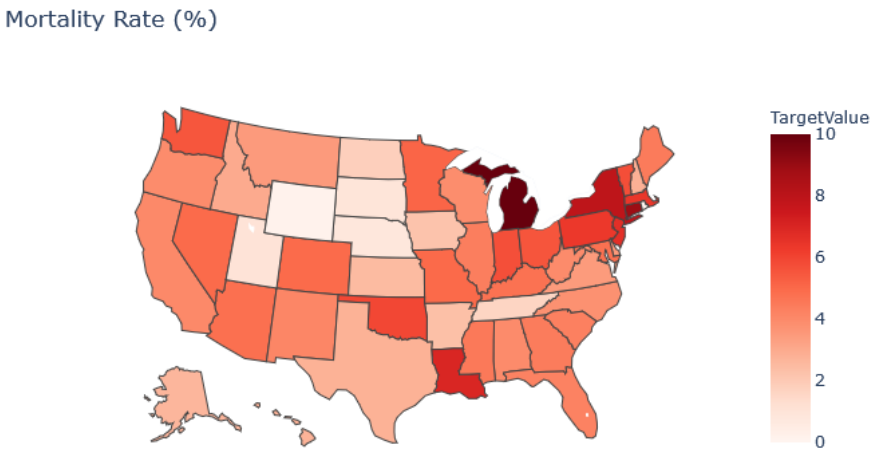

Figure 10: State Mortality Rates

The apparent variation across countries is likely somewaht skewed by differences in testing capacity and reporting accuracy. The variation across U.S. states should be less affected by these factors. The significant variation in mortality rates across states seemingly indicates that mortalitly rates are significantly affected by factors not accounted for in this dataset. It would be useful to gather more detailed data to investigate what might be driving the disparity in mortality rates across U.S. states.

Challenges

I encountered a few challenges in this project. One challenge was the small number of features in this dataset and the absence of some potentially very informative features, such as regional ages. Another challenge was working with time series data, which I had never trained a mode for prior to this project. Because of that, I was uncertain how to approach this problem.

Approach

Data wrangling and feature engineering

I did relatively little feature engineering in this project. I added a day of the week features to try to capture possible variation between weekday and weekend reporting. I converted the date features to integer format to be used in training my model.

One thing that was important to visually exploring the data was not double counting data. The U.S. data seemed to be reported twice, once for the country as a whole and once on per-county basis. This was something I considered in manipulating the data to produce visualizations.

Model selection

I implemented two sklearn regressors–RandomForestRegressor and ExtraTreesRegressor. I split my oringinal training data into approximately 85% training (from the first 85% of dates) and 15% validation (from the last 15% of dates). I noticed several Kaggle notebooks did not take this approach in splitting training and validation sets, but shuffling the data and then splitting instead. It seems wrong to train on data from future dates to validate by prediction on past dates, but perhaps that is a misconseption on my part. Nonetheless, I used training data that preceded the data I used to make validation predictions. Below are plots of validation loss for the two methods I used. Note that I predicted cases (1) and fatalitites (2) separately. I did not find the regression models to be sensitive to the parameters I modified. This is something that should be explored further.

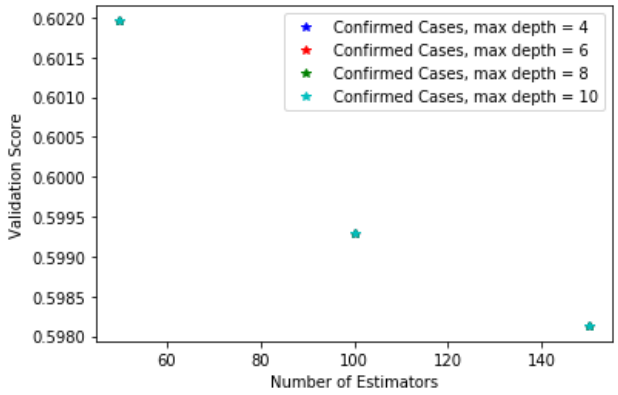

Figure 11: Random Forest Validation Loss for Cases Predictions

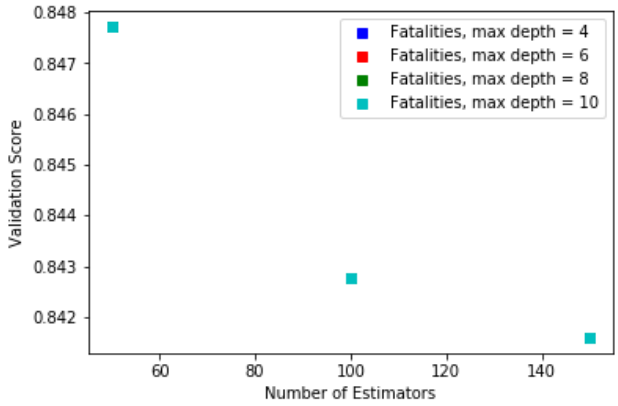

Figure 12: Random Forest Validation Loss for Fatalities Predictions

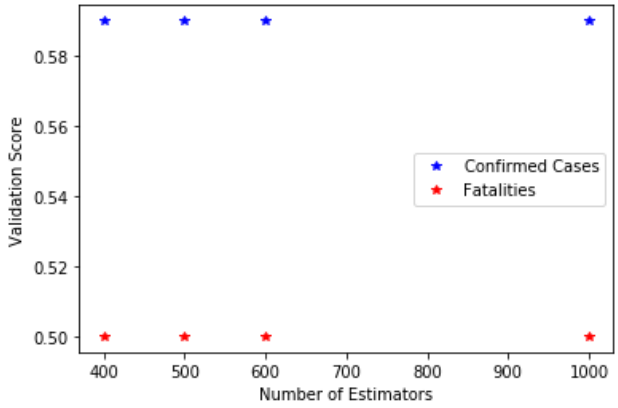

Figure 13: Extra Trees Validation Loss for Cases and Fatalities Predictions

This Extra Trees regression model showed the best validation performance, and so I used that model to make predictions on the test data.

Evaluation and Summary

My model resulted in loss = 0.54 for daily cases predictions and loss = 0.41 for daily fatalities predictions. Shown below are comparisons of my predicted daily cases and fatalities to actual daily cases and fatalities. To visualize the results, I compared the aggregated predictions per data to the aggregated recorded values per date.

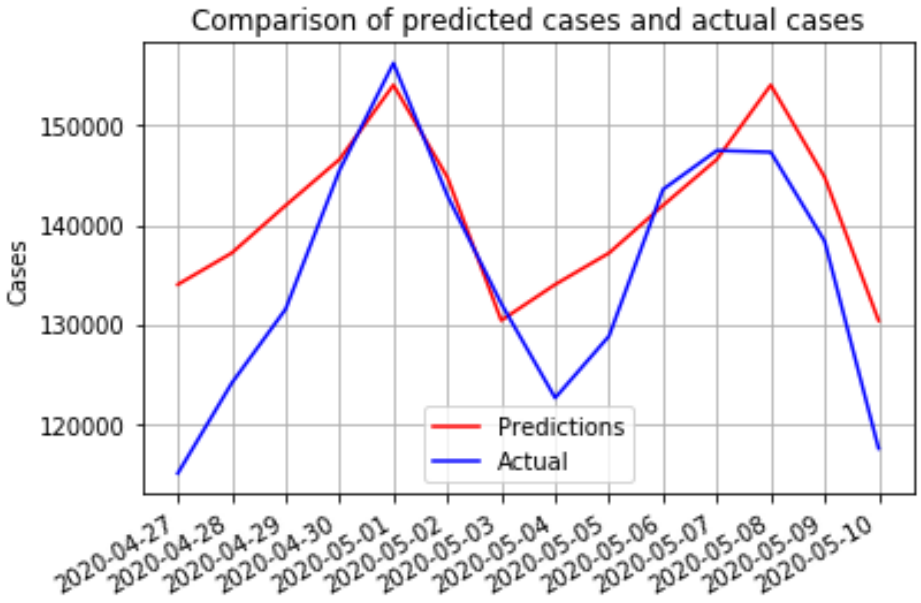

Figure 14: Predicted Daily Cases and Recorded Daily Cases

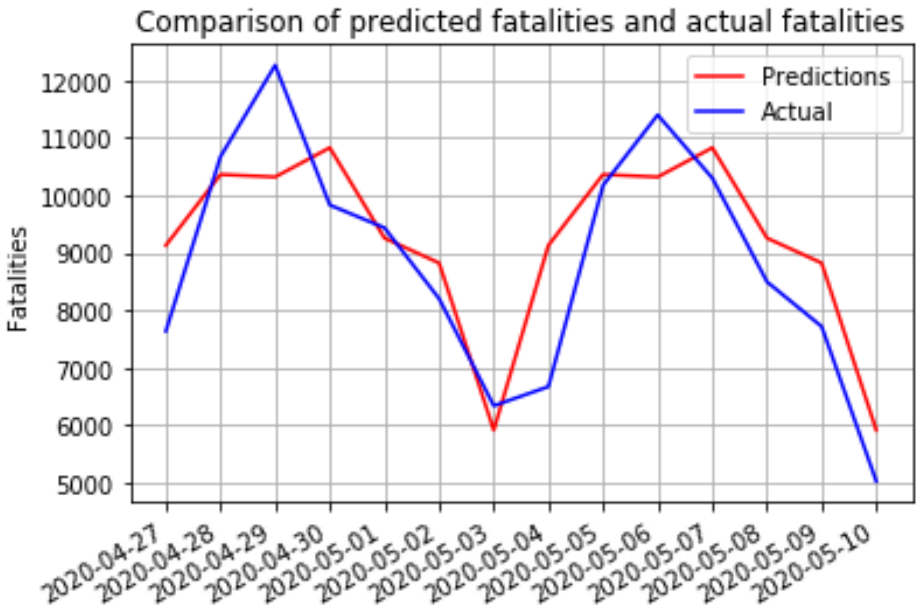

Figure 15: Predicted Daily Fatalities and Recorded Daily Fatalities

My model preformed better in predicting daily fatalities than in predicting daily cases. The recorded data exhibit some sharp and erratic discontinuities, which my model did not capture well. Considering the widespread nature of COVID19 predictions that have received a lot of attention recently, perhaps this model does not perform too poorly. Certainly, though, there is room for improvment. Two factors that could be explored to improve predictions are pulling in more data features and searching for a better approach to

modeling this time-dependent problem.

What I Learned

The main skills I learned through this project were various data visualization techniques. I had not created most of the types of plots used in the project previously. Also, I gained more experience working with data in Pandas and learned some very useful new methods. I'm not sure that I would consider it a skill gained yet, but this was the first time I have applied machine learning to a time-dependent problem. This is an area that I would like to understand better.

References

[1] (2020) Covid-19 forecasts. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

[2] (2020) Covid-19: current situation on may. [Online]. Available: https://www.kaggle.com/corochann/covid-19-current-situation-on-may