Michael Abramson
Santiago Romero
Ye Liu
Junjie Huang

Lumiata Covid-19 Global Hackathon Submission Write-Up

Introduction:

The purpose of this write-up is to provide a detailed description of our team's submission for Lumiata's Covid-19 Global Hackathon. The submission includes this write-up, a video presentation, and a link to our GitHub repository containing all code and datasets. Our submission is a machine learning regression model for predicting spread rates in any given US state. Identifying this rate as our target for study was the first step in our process. This write-up will be divided into sections based on steps in our methodology.

Thought Experiment:

After identifying the differences in rates of COVID-19 cases spread from state to state as our target of study, we conducted a thought experiment to hypothesize what variables might impact this rate. We also took into consideration that these variables have to be accessible or engineerable from existing datasets. Ultimately, we decided that our variables of interest were financial, commute, population data, and COVID-19 case data.

Locating Data Sources:

We proceeded to locate data sources for our variables of interest. These sources include the U.S. Census Bureau, the U.S. Department of Commerce, the 2013 American Community Survey, usgovernmentspending.com, worldpopulationreview.com, and the John Hopkins Covid-19 dataset.

Cleaning and Preparing Data:

We had to engineer our target variable of spread rate from the John Hopkins Covid-19 dataset. We did this grouping and summing total cases per day into states. We then subset these data frames based on the condition that there were more than 10 cases and cases were increasing each day. We used this rules to define whether or not an outbreak was occuring. We then averaged the percentage increases from day to day into a new data frame by state. Lastly, we joined this dataframe with financial, commute, and population data from our other sources.

Modeling:

For our machine learning regression model, we used an ensemble of hyper-parameter tuned gradient boosted regressors and K nearest neighbor regressors. We achieved a final R^2 on the test data of 67.9%, and a final R^2 on the training set of 99%. Unfortunately, this model is admittedly somewhat overfit, but the dataset we worked with is very small, as there are only 50

states, so it is difficult not to overfit a model given this size. Upon completing our model, we examined feature importances which revealed the percentage of commuters using public transportation and average household size to be the most important determinants of spread rate in any given state.

Research:

Once we identified these primary determinants in our model, we conducted research to understand why these two variables are so impactful. We found that public transportation allows for the virus to spread easily across large groups of people. Additionally, we found that a large portion of cases are the direct result of family members spreading the virus to each other within their households.