Project 4: Covid 19 Forecasting

Summary: The first case of Coronavirus (COVID-19) was confirmed in Wuhan, China, in early December, 2020. The virus has since spread across the globe, with over 4 million confirmed cases and 1.4 million deaths.[3] This has created a demand for empirical models to forecasting future COVID-19 cases and fatalities. Such models would allow policy makers and health care providers to more effectively plan for the future. In this report, I use ensemble machine learning techniques to build predictive models for the number of COVID-19 cases and fatalities in the next year. The models, trained on features engineered from daily reports of COVID-19 cases and fatalities, perform well and demonstrate the potential of predictive models on the global stage. This report also investigates a new research questions — in the US, what state-level demographic and economic covariates are the most informative predictors of a state's COVID-19 mortality rate? I find that educational attainment and the fraction of the labor force employed in the service industry are the most informative predictors, while age structure is surprisingly less informative. In tandem, these two analyses demonstrate that (1) statistical models can effectively forecast future COVID-19 cases and deaths and (2) COVID-19 forecasting can be improved by including state-level demographic and economic measures.

Data Processing: The data for the forecasts were obtained from the COVID-19 Global Forecasting project. The data includes variables reporting the count of COVID-19 confirmed cases and fatalities at the national level. For a few larger countries, data was also available at the sub-national level. The data is available at a daily time-stamp from January 22nd, 2020 to April 5th, 2020. The data was pre-processed, and I performed no data cleaning. I split the original data into a training dataset with observations from January 22nd, 2020 to March 20th, 2020 (17,136) and a test dataset with observations from March 20th, 2020 - April 5th, 2020) (N = 5,814).

Feature Engineering: My exploratory data analysis (EDA) revealed a pattern where after the first COVID-19 case was reported in a region, that region experienced exponential growth in the total number of reported cases. As demonstrated in Figure 1, there was significant heterogeneity by region on date of first confirmed COVID-19 case. These figures match our intuition for a disease with an R_0 greater than 0; once we observe the first case of COVID-19 in a region, we would expect counts of confirmed cases to grow exponentially. The temporal heterogeneity highlights the importance of accounting for the temporal aspect of COVID-19 in any forecast. For feature engineering, I constructed two variables (i) days since first confirmed case and (ii) days since first fatality. A value of 0 for these variables denotes there haven't been any confirmed cases.

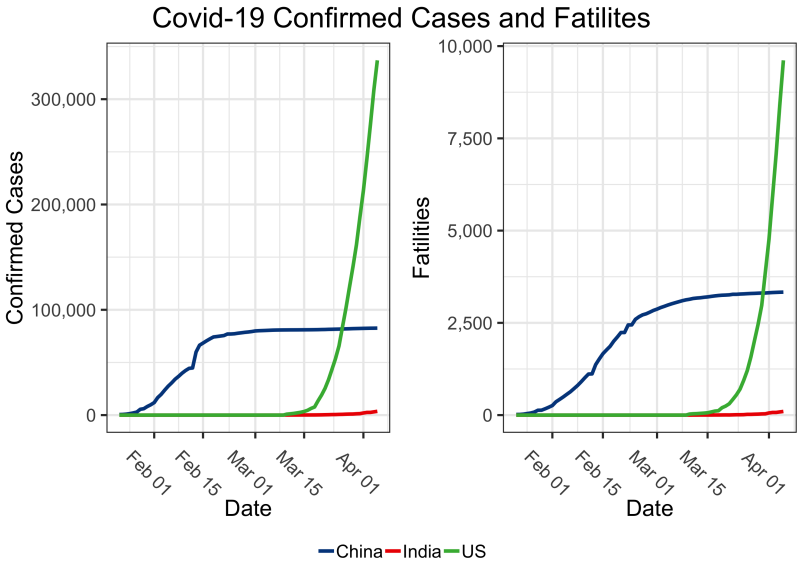


Figure 1: Estimates for China, India, and the US.

RMSE	GBM	GLM
Confirmed Cases	1.96	2.14
Fatilities	2.97	4.64

Modeling: I fit two separate predictive models to estimate the number of COVID-19 cases and fatalities: a generalized linear model (GLM) and a gradient boosted model (GBM). The GLM model is a generalization of a linear regression, which allows the errors to be have distributions other than the normal distribution. The GBM model fits a series "weak" models and ensembles them together into a more powerful "committee," allowing for accurate prediction. The dependent variable for both models was either (i) counts of confirmed COVID-19 cases or (ii) count of COVID-19 fatalities. The predictor variables were (i) geography, at either the country or the sub-national level, if available and (ii) days since first confirmed case or days since the first confirmed fatality (respective to the dependent variable). These algorithms were fit using 10-fold cross validation, where the sample was split into 10 folds. For each of the 10 folds, I hold that fold out as the test dataset. I then fit a model using the other 9 folds at the training dataset. I repeat this process for every one of the 10 folds. The optimal combination of all models is then used for the final algorithm.

The performance of the models is measured using the root mean squared logarithmic error (RMSLE), the same metric employed by the original Kaggle competition: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1)-log(a_i+1))^2}$, where p_i is predicted number of confirmed cases (or fatalities) and a_i is actual number of confirmed cases (or fatalities), and n is the number of observations in the test dataset. The table above shows the RMSE for the two models.

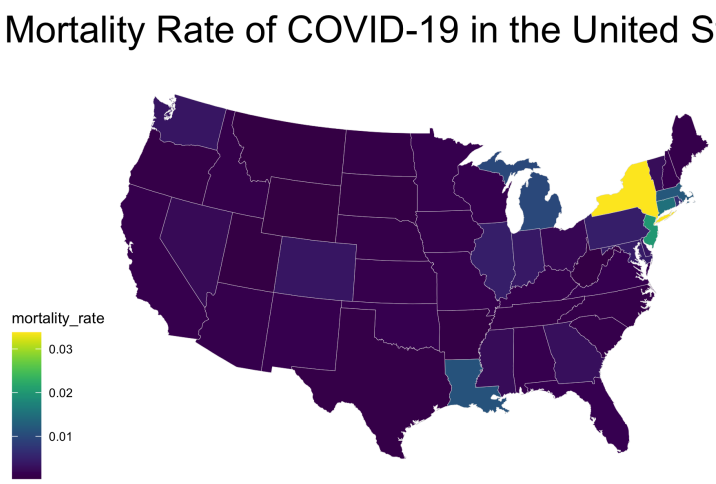


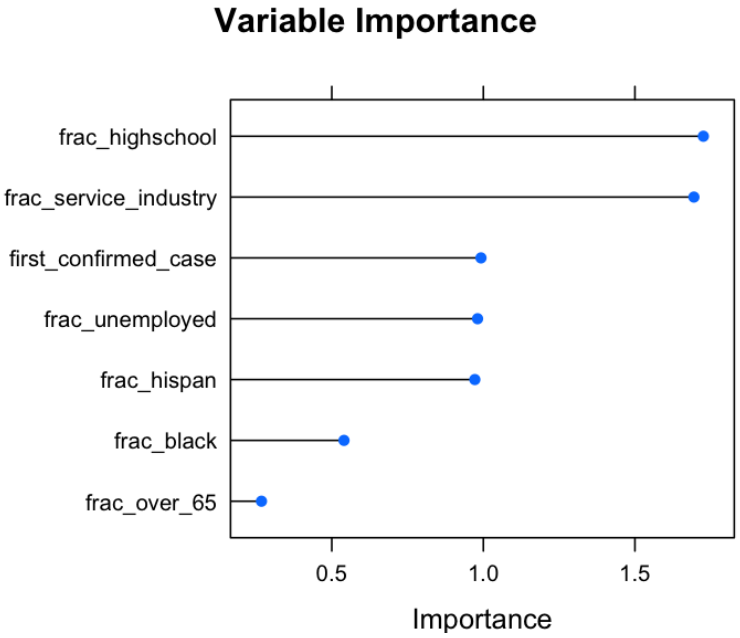
Figure 2: Geographic Variation in Covid-19 Mortality Rates

Original Question: Several research teams have shown the relationship between the age structure of a geographic area and the COVID-19 mortality rate.[4] Several other demographic covariates, such as educational attainment and fraction of the labor force employed in a service-industry job are hypothesize to be powerful predictors of the mortality rates of COVID-19, yet the association between these covariates and COVID-19 mortality rates remains an open topic of research. This inspires my original research question — which state-level demographic and economic characteristics are the strongest predictors of the state-level mortality rate from COVID-19? I define the mortality rate of COVID-19 as $M_i = \frac{\text{Total COVID-19 Fatilities in state } i}{\text{Total Population in State } i}$. Figure 2 shows the significant state-level variation in mortality rates for COVID-19, the motivation for this analysis. New York, Washington, and Louisiana have notably high mortality rates — is this noise, or is there some signal that can be explained by the demographic composition of these states?

To answer this question, I obtain two new datasets: (1) state-level COVID-19 death counts from the New York Time COVID-19 repository and (2) the American Community Survey (ACS) 2019 survey, which includes relevant demographic and economic covariates. I first link the data sets at the state level and calculate the total COVID-19 mortality rate for each state. I then construct a

set of 7 different COVID-19 state-level covariates using the ACS data: proportion Black, proportion Hispanic, proportion graduating from high school, proportion working in the service industry, proportion unemployed, and proportion over 18. I select these state-level COVID-19 covariates based off of demographic domain knowledge and the work of Dowd et al. 2020. Given the superior performance on the GBM model in my forecasting analysis, I fit a model to calculate variable importance. As I am interested only in the explanatory power of the variables, I do not break the dataset into a training and test data. Instead, I run my analysis on all 50 states. I calculate variable importance (relative influence) as the the average variable importance across all trees generated by the boosting algorithm[1].

Figure 3 shows the results of out variable importance calculations. Education matters — it was the strongest predictor of mortality rate. Additionally, the fraction of the population working in the service industry was an important predictor of COVID-19 mortality rates. Notably, the fraction of the population over the age of 65 was not a strong predictor of mortality rates. Identifying the predictors most strongly associated with COVID-19 facilitates building more better models in the future.



Discussion

In tandem, these analyses present a promising step forwards for COVID-19 forecasting. While forecasting infectious diseases is challenging, our models produced accurate predictions despite a limited set of covariates. This suggests that we can effectively forecast COVID-19 with models, but that we must be careful to account for outliers. Future work could investigate the regions that are performing better than our model predicts, and use these countries as the "positive exemplars" on which to base future analysis. My original research question shines light on the predictors of mortality rates in the United States. By fitting a GBM algorithm with covariates obtained from the ACS, I identified the most influential predictors of a state's COVID-19 mortality rate. This compliments our previous analysis by incorporating important social factors. While forecasting COVID-19 cases and fatilities is feasible with counts alone, incorporating the demographic and economic covariates can further improve our model's predictive power.

Citations: [1] COVID-19 Global Forecasting. Kaggle Competition. May 2020. [2] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 - 26. [3] WHO report on Coronavirus. World Health Organization. Retrieved 4 May 2020. [4] Dowd, J. B. et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. OSF https://osf.io/fd4rh (2020).

[1]This calculation was carried out using the `varImp` function in Caret (Kuhn 2008)