Modeling and Predicting COVID-19 Infection Trends

TR1ND1

nrshaida, dcheong, rmani1, tphilli6

Goal

The COVID-19 pandemic has, undoubtedly, had a tremendous impact all across the globe, with the ramifications of the outbreak likely to last well into the future. Without knowledge of how infection numbers will change over time, it is significantly more difficult to propose prophylactic and deterrent measures as well as plan for the reopening of society. Our goal in this project is to forecast COVID-19 infection trends up to a week in advance for locations affected by COVID-19, given relevant data including current infections, susceptible population size, recovered cases, deaths, and social mobility.

Data

Data pertaining to infection, recovery, and death numbers were procured from Johns Hopkins University's COVID-19 data repository. The repository contains global infection data since January 23rd, with updates taking place frequently each day henceforth. Although this is realistically not the case, for the sake of this project we are assuming that each country's reported numbers are accurate. Population data was obtained from the United Nations. Government mitigation data was compiled from ACAP, an independent information provider that aggregated data for 192 countries since January 1st. Social mobility data (based on smartphone use) from March 1-April 9 was obtained from the geospatial analysis company Descartes Lab.

Model+Evaluation Setup

To forecast COVID-19 infect trends, we made use of both a linear/MLP regression model and a time-series model. In the former, we selected the five most correlated features (infected, recovered, previously infected, deaths, and days since the onset of the outbreak) and predicted next-day infections, recovered, and deaths using a linear regressor with lasso regularization and normalized variables as well as a multi-layer perceptron regressor with early stopping. These predictions were fed back into the model to forecast up to seven days in advance. A randomly distributed train/test split of 8:2 on global locations was utilized with completely held-out test data. In the latter model, a smaller-scale analysis was performed on the 50 U.S. states utilizing data external to the virus, specifically movement of people (social mobility). Training was done on the concatenation of 40 random states, with testing done on those remaining (forecasting a week in advance). Our sole parameter (number of lags) was chosen to be 8 days due to the delay between infection and infection detection being approximately this length. The model was run on 3 different ways of measuring infection rates: Pure VAR, Percent VAR, and Log VAR.

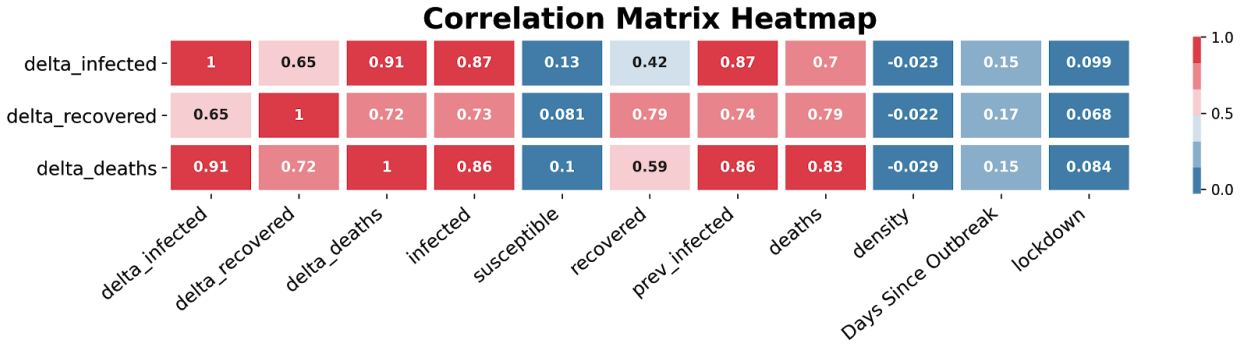
Results and Analysis

Claim #1: Our models underperform compared to the baseline models.

Support for Claim #1: As seen in the figure below our baseline predictors mean positive, negative, and percentage error is less than that of our linear regressor, MLP regressor, and our time series predictor.

Error Metric*\Predictor	Baseline predictor (last-known moving average of new infections, recovered, deaths)	Linear Regressor	MLP regressor	Baseline (last-known rate of new infections)	Time Series 1 (Pure)	Time Series 2 (Percent)	Time Series 3 (Log)
Mean positive error (infections)	107.2	607.2	309.7	242	3137	23639	2948
Mean negative error (infections)	-129.0	-322.8	-241.9	-174	0	0	0
Mean percentage error	29.1%	13112.0%	694.5%	0.3%	164%	530%	100%
*Error metrics are calculated as the mean error across all possible 7-day forecasts for all testing locations in the testing datasets.							

Claim #2: Most of the predictive power of our linear regressor/MLP regressor comes from infected, recovered, previous infected, deaths, and days since outbreak.



Support for Claim #2: From the figure above, we were able to determine that the most significant features to include would be infected, recovered, previous infected, deaths, and days since outbreak. Susceptible infected was also correlated with delta_infected, but we suspected that this could be explained by the target variable's correlation with infection numbers.

Claim #3: From our time-series model of infections and social mobility data, we reject the null hypothesis that mobility does not affect infections.

	R^2 (train)	R^2 (test)	Root MSE (train)	Root MSE (test)	Inf <- Mobility p-value	Inf -> Mobility p-value
Pure VAR (1)	0.93	0.88	3,097	1,125	0.000	0.710
Percent VAR (2)	0.20	0.71	11,384	11,618	0.000	0.000
Log VAR (3)	0.95	0.94	4,086	931	0.000	0.000

Support for Claim #3:

The last two columns indicate the results of a Granger causality test. As the p-values for the null hypothesis that mobility does not affect infections is less than 0.05, we reject the null hypothesis.