An Inferential Analysis of Predictive Modeling for the Spread of COVID-19

Michael Kolor

May 15 2020

1 Abstract

In this paper, the use of predictive modeling is explored in obtaining both an empirical accuracy of United States COVID-19 spread rates and an inferential understanding of why transmission occurs. This paper walks through and analyze a progression of modeling techniques, starting with less powerful but more inferential models such as lasso regression and advancing towards more powerful, higher-level models such as Long Short-Term Memory (LSTM) Neural Networks. Data for the models was aggregated across a vast array of sources to explore many different factors for COVID-19. The results of the modeling process were limited by the complex nature of the problem, but supported the presence of regionalization in COVID-19 spread and higher-prevalence in urban areas, along with a sequential relationship for COVID-19 spread. Individual models also supported the prevalence of a delayed effect in certain policies with the prevalence of COVID-19 spread.

2 Introduction

The world is currently facing a pandemic of the coronavirus disease 2019, or COVID-19. Within 100 days of the first reported cases, the pandemic had already infected millions, forcing draconian shutdown measures around the globe. The virus' ability to be transmitted asymptotically and the lack of human immune resistance for it has been a major limitation in attempts to contain the spread. In the United States, there are over 1.4 million currently confirmed cases. However, policymakers across the country have begun implement measures to reopen society on a regional basis. Such policymakers are in strong need of data and information on the spread of disease to assist in current and future decisions regarding the matter. Many sources are being explored to glean information about the pandemic, and much data is available to the public [3].

3 Research Question

The research question for this paper is what information can be learned about the spread of COVID-19 through the use of predictive modeling. The response variable that these models will focus on is predicting the amount of new cases per person that are confirmed on a certain day, grouped by county.

4 Methods

4.1 Data Collection

To facilitate the needs of the modeling and incorporate as many inferential opportunities as possible, modeling data has been collected, aggregated, and pre-processed across numerous different resources. This was all done in a Python environment.
Sources used to collect raw data were Johns Hopkins University, the Malaria Atlas Project, StreetLight, the IMHE, and Kaggle. The dates in focus for this project were March 1 to April 17.

4.2 Feature Engineering

To gather a metric for proximity to urban areas, a geospatial map of travel time in minutes to the nearest urban area was averaged across each county's area in GIS. For necessary models, the states that each county is located in were one-hot encoded. Air Quality Index (AQI) levels for four major pollutants for each county were listed, and imputed across state or national averages to handle missingness. To approximate movements, Vehicle Miles Traveled (VMT) measurements from the company StreetLight were used as a feature as well as the January VMT as a baseline. Demographic and mortality-related features were also added. To create a feature measuring policy response to COVID-19 in an area, a dataset storing summarized policies implemented in the United States was aggregated over four different categories and summed by day to create features representing the number of policies implemented in that county's state on a given day by policy type.
Upon review of scatterplots between features and the raw response variable, total new cases in that county on that day, it appeared that many data transformations needed to be performed. This was further modeling showed heavy bias towards *population* as a feature. In response, the problem was adjusted to a "per person" approach to remove population weighting and gain more interesting insights into the COVID-19 spread. Such features that needed to be altered to per-person were done so. Still, some features needed transformed. Taking the square root of the urban accessibility feature generated a more normal distribution and more linear relationship with the chosen response metric, while converting the per-person movement features to their logarithmic value (with 1 added to handle edge cases) was preferred. For all features involving case counts, converting their metric to the logarithmic value of the

number of cases per 100,000 people corrected distribution and relationship non-linearity.

Table One of the Design Matrix Data

	count	mean	std	min	25%	50%	75%	max
log_new_cases_per_100k	140248.0	0.396949	0.835410	0.000000	0.000000	0.000000	0.000000	6.691202
median_age	140248.0	41.101173	5.157152	21.700000	38.000000	41.100000	44.100000	67.000000
female_percentage	140248.0	49.984532	2.313368	21.003945	49.500719	50.424618	51.127277	57.922691
sqrt_city_accessibility	140248.0	8.089619	3.209171	0.000000	5.945310	7.909350	9.917718	19.237817
NO2 AQI	140248.0	20.302064	6.087179	1.542023	15.866779	20.210284	24.479102	51.903226
O3 AQI	140248.0	37.214672	4.906582	22.618388	34.486920	36.226616	40.347497	60.124481
SO2 AQI	140248.0	7.251697	5.523718	0.077598	2.924696	6.415778	10.369300	28.976401
CO AQI	140248.0	4.670114	1.662533	0.000000	3.619607	4.396789	5.567241	16.074074
log_january_avg_vmt_per_person	140248.0	3.705117	0.503166	1.021683	3.425649	3.730354	4.007360	8.771737
log_county_avg_vmt_per_person	140248.0	3.242546	0.755673	0.220534	2.750069	3.265932	3.751983	9.421730
lower_respiratory_infection_and_other_infection_mortality	140248.0	33.719910	10.424626	9.080000	26.625000	32.570000	39.472500	90.830000
chronic_respiratory_disease_mortality	140248.0	64.298063	16.772486	14.270000	52.540000	62.960000	73.715000	160.970000
mortality	140248.0	880.972875	142.257566	309.360000	781.170000	868.315000	975.092500	1570.450000
new_travel_measures	140248.0	0.002439	0.049321	0.000000	0.000000	0.000000	0.000000	1.000000
new_school_measures	140248.0	0.021861	0.160770	0.000000	0.000000	0.000000	0.000000	2.000000
new_soc_dist_measures	140248.0	0.110511	0.492474	0.000000	0.000000	0.000000	0.000000	7.000000
new_state_of_emergency_measures	140248.0	0.009312	0.096049	0.000000	0.000000	0.000000	0.000000	1.000000
new_testing_measure_measures	140248.0	0.043858	0.204780	0.000000	0.000000	0.000000	0.000000	1.000000
log_new_cases_per_100k_prev_1_days	140248.0	4.713237	72.480811	0.000000	0.000000	0.000000	0.000000	7837.000000
log_new_deaths_per_100k_prev_1_days	140248.0	2.350836	98.058280	0.000000	0.000000	0.000000	0.000000	13374.000000
log_new_cases_per_100k_prev_5_days	140248.0	0.457436	0.757694	0.000000	0.000000	0.000000	0.778440	5.779512
log_new_deaths_per_100k_prev_5_days	140248.0	0.139647	0.476758	0.000000	0.000000	0.000000	0.000000	6.620867
log_new_cases_per_100k_prev_7_days	140248.0	0.451164	0.735563	0.000000	0.000000	0.000000	0.754687	5.769571
log_new_deaths_per_100k_prev_7_days	140248.0	0.130063	0.453525	0.000000	0.000000	0.000000	0.000000	6.564784
log_new_cases_per_100k_prev_14_days	140248.0	0.398986	0.661295	0.000000	0.000000	0.000000	0.634046	5.746626
log_new_deaths_per_100k_prev_14_days	140248.0	0.102291	0.380211	0.000000	0.000000	0.000000	0.000000	6.324047
new_travel_measure_0_7_days_ago	140248.0	0.038346	0.192032	0.000000	0.000000	0.000000	0.000000	1.000000
new_school_measure_0_7_days_ago	140248.0	0.153029	0.525245	0.000000	0.000000	0.000000	0.000000	5.000000
new_soc_dist_measure_0_7_days_ago	140248.0	0.773180	1.607996	0.000000	0.000000	0.000000	1.000000	17.000000
new_state_em_measure_0_7_days_ago	140248.0	0.069406	0.254144	0.000000	0.000000	0.000000	0.000000	1.000000
new_testing_measure_0_7_days_ago	140248.0	0.307006	0.561950	0.000000	0.000000	0.000000	1.000000	3.000000
new_travel_measure_7_14_days_ago	140248.0	0.166006	0.372087	0.000000	0.000000	0.000000	0.000000	1.000000
new_school_measure_7_14_days_ago	140248.0	0.153029	0.525245	0.000000	0.000000	0.000000	0.000000	5.000000
new_soc_dist_measure_7_14_days_ago	140248.0	0.774877	1.607706	0.000000	0.000000	0.000000	1.000000	17.000000
new_state_em_measure_7_14_days_ago	140248.0	0.070425	0.255863	0.000000	0.000000	0.000000	0.000000	1.000000
new_testing_measure_7_14_days_ago	140248.0	0.307006	0.561950	0.000000	0.000000	0.000000	1.000000	3.000000
new_travel_measure_14_21_days_ago	140248.0	0.164323	0.370570	0.000000	0.000000	0.000000	0.000000	1.000000
new_school_measure_14_21_days_ago	140248.0	0.152958	0.525198	0.000000	0.000000	0.000000	0.000000	5.000000
new_soc_dist_measure_14_21_days_ago	140248.0	0.773993	1.607857	0.000000	0.000000	0.000000	1.000000	17.000000
new_state_em_measure_14_21_days_ago	140248.0	0.070425	0.255863	0.000000	0.000000	0.000000	0.000000	1.000000
new_testing_measure_14_21_days_ago	140248.0	0.307006	0.561950	0.000000	0.000000	0.000000	1.000000	3.000000
state_Alabama	140248.0	0.022453	0.148152	0.000000	0.000000	0.000000	0.000000	1.000000

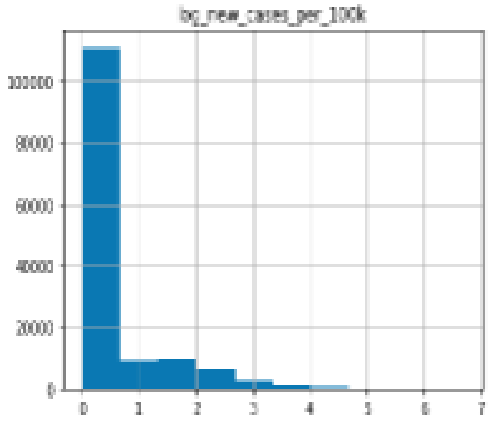
For more plots of the distributions of the features, their correlation, and their relationship with the predictor, see Appendix A, B, C.

4.3 Response

Similar to the case count features, the response variable was transformed from case counts per day to the log of the case counts per 100,000 people, creating a more normal distribution of values. However, the distribution was still very skewed towards 0. Below are summary visuals:

```
Distribution of Log (New_Cases per 100,000 People + 1)

count    140248.000000
mean          0.396949
std           0.835418
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           6.691202
Name: log_new_cases_per_100k, dtype: float64
```



4.4 Managing Lag Time

Inherently, the occurence of new COVID-19 cases is a sequential problem. There is a dependency on the number of COVID-19 cases in previous days to helpe predict the number of COVID-19 cases on a given day. However, certain models are not capable of handling such stochastic processes. As an approximation, features were added that incorporated a rolling average in new cases and new deaths per 100,000 people over a certain time period. Background literature on the incubation period is limited and indicates a wide range of 2-14 days before symptoms are shown, and up to a five day period before tests returned positive [1, 4]. To explore this relationship, multiple past windows of time for COVID-19 case counts were added as features to the design matrix, as well as for deaths, which may be indicators of higher prevalence of COVID-19 in a community than tests are showing [3]. The same process of windowing past counts was also applied to features regarding the number of enacted policies by a state in hopes to best determine the amount of time to be expected before effects of certain COVID-19 policies can be seen in COVID-19 case counts [1].

A problem that can result from the addition of these windowed features is multicollinearity (Appendix A), which can hinder performance for certain models. This problem was considered in the selection of initial models, result-

ing in the use of L1 regularization [8].

4.5 Modeling

Several modeling approaches were utilized sequentially to predict COVID-19 cases counts and make inferential conclusions.

4.5.1 Lasso Regression

The first model utilized was a regularized version of an Ordinary Least Squares Regression known as a Lasso Regression. This model was created using the *statsmodels* in Python. Lasso regression involves modify the linear regression loss function by adding a penalty term that is the sum of the absolute value of all coefficients in a model, multiplied by a constant, α (which was tuned to 0.01). This is illustrated below: [8]

Model: Loss Function + Penalty Term

Squared Loss Sum of Absolute Weights

$$\text{argmin}_\beta \sum_i (y_i - \sum_j x_{ij}\beta_j)^2 \quad +\lambda \sum_j |\beta_j| \qquad \ell_1 \text{ Penalty}$$

Lasso

Such a loss function encourages sparsity, meaning that it favors shrinking coefficients of less relevant features to 0's. This can make lasso regression useful for feature selection and elimination of correlated features, which is necessary to handle the 91 features in the design matrix. Lasso regression still makes the same assumptions as unregularized linear regression, which is that all features are independent of each other and linearly related to the response variable [8].
Lasso regression is computationally efficient and much easier to interpret than many sophisticated models, making it a natural starting point for inferential understanding [7].
The following features comprised the design matrix for the lasso regression:

```
Columns used in design matrix:
        median_age
        female_percentage
        sqrt_city_accessibility
        NO2 AQI
        O3 AQI
        SO2 AQI
        CO AQI
        log_january_avg_vmt_per_person
        log_county_avg_vmt_per_person
        lower_respiratory_infection_and_other_infection_mortality
        chronic_respiratory_disease_mortality
        mortality
        new_travel_measures
        new_school_measures
        new_soc_dist_measures
        new_state_of_emergency_measures
        new_testing_measure_measures
        log_new_cases_per_100k_prev_1_days
        log_new_deaths_per_100k_prev_1_days
        log_new_cases_per_100k_prev_5_days
        log_new_deaths_per_100k_prev_5_days
        log_new_cases_per_100k_prev_7_days
        log_new_deaths_per_100k_prev_7_days
        log_new_cases_per_100k_prev_14_days
        log_new_deaths_per_100k_prev_14_days
        new_travel_measure_0_7_days_ago
        new_school_measure_0_7_days_ago
        new_soc_dist_measure_0_7_days_ago
        new_state_em_measure_0_7_days_ago
        new_testing_measure_0_7_days_ago
        new_travel_measure_7_14_days_ago
        new_school_measure_7_14_days_ago
        new_soc_dist_measure_7_14_days_ago
        new_state_em_measure_7_14_days_ago
        new_testing_measure_7_14_days_ago
        new_travel_measure_14_21_days_ago
        new_school_measure_14_21_days_ago
        new_soc_dist_measure_14_21_days_ago
        new_state_em_measure_14_21_days_ago
        new_testing_measure_14_21_days_ago
        state_Alabama
        ...
```

4.5.2 Linear Mixed-Effects Model

The second model utilized for inference was the linear mixed model. Linear mixed models are an extension of simple linear models to allow both fixed and random effects. They are used when to manage nonindependence in the data that results from a hierarchical structure. This is useful to explore a hierarchical nature of the given data; counties could be considered as grouped within states, as states could have varying governmental, geographic, and demographic differences. The specific mixed model that is utilized is a varying coefficients, or random slopes, model, with the equation listed below:

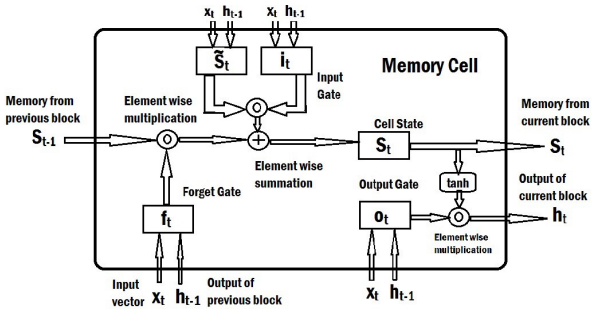
$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_{0i} + \gamma_{1i} X_{ij} + +ij$$

Above, Y_{ij} is the j-th measured response for a county , and X_{ij} is a covariate for this response. The "fixed effects parameters" β_0 and β_1 are shared by all states, and the errors ϵ_{ij} are independent of everything else, and identically distributed (with mean zero). The "random effects parameters" γ_{0i} and γ_{1i} follow a bivariate distribution with mean zero, described by three parameters: $\text{Var}(\gamma_{0i})$, $\text{Var}(\gamma_{1i})$, and $\text{cov}(\gamma_{0i}, \gamma_{1i})$. There is also a parameter for $\text{var}(\epsilon_{ij})$. This model is fit to maximize log-likelihood, without an analytic solution. The objective function for determining log-likelihood can vary, and for this instance the Python package *statsmodels*' implementation is used. Because of its computational complexity, this model was limited only to parameters found significant in the lasso regression model, grouped by state [5].

4.5.3 Long Short-Term Memory Network

While lasso regression and mixed-effects modeling can provide very interpretive insight on the features, they are not well-equipped to handle the inherent sequential nature of the problem at hand as well as the interactions and non-linear relationships between features and the response variable as well as

between the features themselves. To account for this, an LSTM Network can be used. LSTM Networks are a special case of a Recurrent Neural Network (RNN), which is a neural network that handles sequential data through utilizing loops in its architecture. Simple RNN's can struggle to handle long-term dependencies in memory, which is why LSTM Networks are used. A visual representation of an LSTM architecture can be seen below [6]:



LSTMs build on RNN's through the use of four interacting layers in the network. These layers include a "forget gate" layer, which, as a gate, controls what memory is forgotten or retained. Another layer is an "input gate", which decides how values will be updated from the newly inputted features. This input is then transformed using a tanh layer to become "candidates" for information to be stored, the decision for which is left to a final "output gate" layer. The model's weights are not determined analytically, but through backpropagation [6].
The design matrix for this model varied from the other models in that inputs were handled in batches sorted by county and date, allowing for the removal of windowed data as the LSTM determines the lag effect inherently. To determine the influence of certain features in predicting COVID-19 cases, subsets of features were removed per model, and the model's final loss was compared to a baseline model including all features.

5 Discussion

5.1 Results

5.1.1 Lasso

The fitted lasso regression model yielded an R^2 Score of 0.470 on the training data and 0.461 on the test data (selected using a 90-10 train-test split). The mean-squared-error was 0.385. The significant predictors as selected by the model were:

```
Significant Predictors from Lasso Regression:
        median_age
        sqrt_city_accessibility
        log_january_avg_vmt_per_person
        log_county_avg_vmt_per_persont
        chronic_respiratory_disease_mortality
        log_new_cases_per_100k_prev_1_days
        log_new_cases_per_100k_prev_5_days
        log_new_deaths_per_100k_prev_5_days
        log_new_cases_per_100k_prev_7_days
        log_new_cases_per_100k_prev_14_days
        new_school_measure_7_14_days_ago
        new_soc_dist_measure_7_14_days_ago
        new_school_measure_14_21_days_ago
        new_soc_dist_measure_14_21_days_ago
        state_Alabama
        state_Kentucky
        state_Louisiana
        state_Massachusetts
        state_Mississippi
        state_New Jersey
```

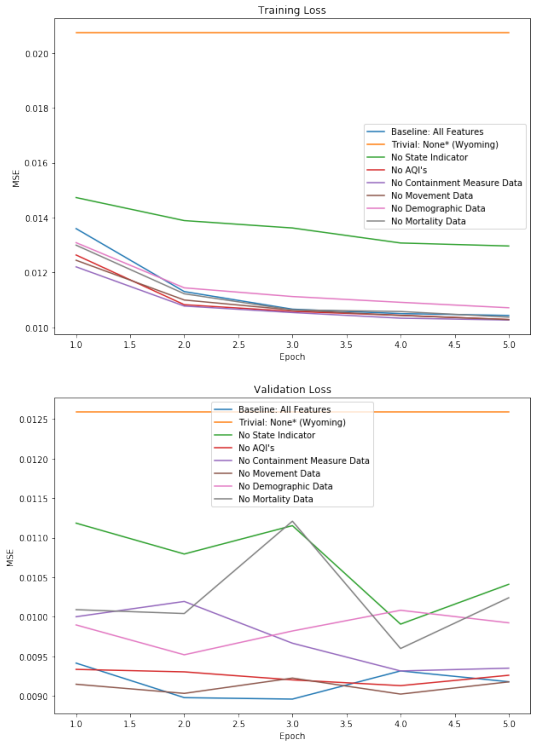
The model was unable to select a specific lag-time as significant in terms of previous case counts, selecting most as significant. However, the model only found school and social distancing measures to be significant, and only between 7-21 days, indicating that those types of policies may have been more significant in affecting COVID-19 spread, but at least a week after implementation. Certain states also had significantly different COVID-19 case rates, according to the model. However, the model seems to show signs of confounding variables, as it indicates that variable *median age* of a county has a negative relationship with COVID-19, which may just be indicative that some rural areas may have older populations. Because of this, results must be treated cautiously.

5.1.2 Mixed-Effects

The fitted fixed-effects model using the selected predictors from the lasso regression, grouped by state, showed very similar results to the lasso regression. The mean-squared-error was 0.35, and all predictors were found to have a significant fixed effect with the exception of *new_school_measures_14_21_days_ago* and *new_school_measures_7_14_days_ago*. This could be due to collinearity, however, the variance of the random slopes was by far the largest for those two features, and also for *new_soc_dist_measures_14_21_days_ago* and *sqrt_city_accessibility*. This seems to indicate a large regional variation by state in the effectiveness of policy by state legislatures in reducing COVID-19 prevalence, and a variation in urban COVID-19 spread by region.

5.1.3 LSTM

Training Performance of LSTM's with Removal of Groups of Features



Baseline Test MSE: 0.0095
Trivial Test MSE: 0.0193
Mortality-Less Test MSE: 0.0096
Movement-Less Test MSE: 0.0093
Demographic-Less Test MSE: 0.0100
AQI-Less Test MSE: 0.0095
Measure-Less Test MSE: 0.0095
State-Less Test MSE: 0.0120

In general, the LSTM models are by far the highest-performing, with mean-squared-errors all below 0.025. Removing the state features led to an increase in loss for the model for both training and test mean-squared-error, as seen above. Also, removing a cluster of demographic features – the median age, female percentage, and city accessibility – seemed to result in a slight decrease in performance. This seems to suggest that these features help to improve the model performance in predicting COVID-19. Because the variances in the model performance were quite small compared to the variance in performance due to different local maxima being found through backpropagation optimization, it was hard to make definitive conclusions on what features aided the model in performance. However, the removal of all features but the response led to significantly worse performance, indicating a complex relationship of these features may help improve performance, which was exploited by the model.

6 Conclusion

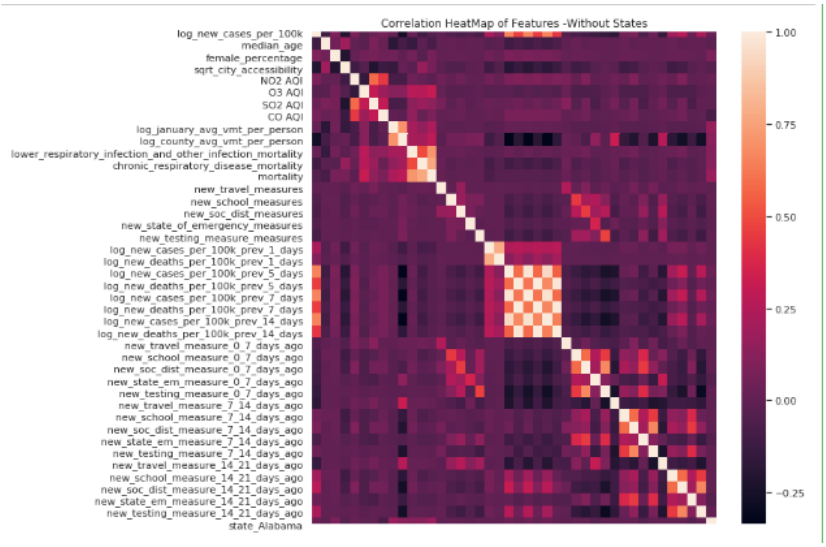
The explored models shared several common trends. They all appeared to support the significance of regional variation in COVID-19 spread, which may support a regionalized approach to handling the pandemic. Previous case counts also remain the vastly important feature for understanding the risk COVID-19 is for a certain county. City-accessibility also appeared to have significance in all models in predicting COVID-19 spread, supporting the idea that clusters for COVID-19 often appear more frequently in urban areas and are harder to control there.
In the exploration of the lag-effect, the lasso and mixed-effects regression appeared to support the idea that effects of school measures and social distancing measures may show significant effects on COVID-19 case counts between 7-21 days after their implementation, supporting the lag times predicted in current researh [1]. The mixed-effects model, however, showed strong regional variance in the effectiveness of such policies.
The LSTM showed an ability to capture much more complex relationships than the other two models, as expected, and supported the idea that the COVID-19 pandemic is a sequential problem. However, its low inferential power made it difficult to make strong conclusions from.

6.1 Limitations/Future Exploration

Exploring the factors for COVID-19 is a complex and limiting topic for many reasons. The high-asymptotic rate makes it difficult to estimate how many true COVID-19 cases exist in a certain region, and testing abilities also vary per region. The transmission is hard to track, as can be seen in policies that have not flatten the curved as expected. Also, there are many confounding variables, such as increases in testing leading to large increases in COVID-19 case counts, leading to misleading relationships being indicated by the models. In addition, it is very early in the pandemic's life cycle in the United States, with only forty-seven days of data available at the time of modeling. Waiting to aggregate more data for modeling yield in much improved results. There is also much incompleteness of data, whether it be the missing tests or missing values for other features, such as the air pollution statistics, which suffered from large missingness [2]. In the case of the used models, the lower-performing models may have showed greater inference, but their poor ability in capturing the complex relationships between features must be acknowledged. Additionally, the higher-performing models captured more complex relationships, but in a manner that was inaccessible for inference. In the future, it may be better to explore types of generalized linear models for solving the current problem inferentially. Because of all of the complex factors involved in this problem, the insights found must be treated judiciously.

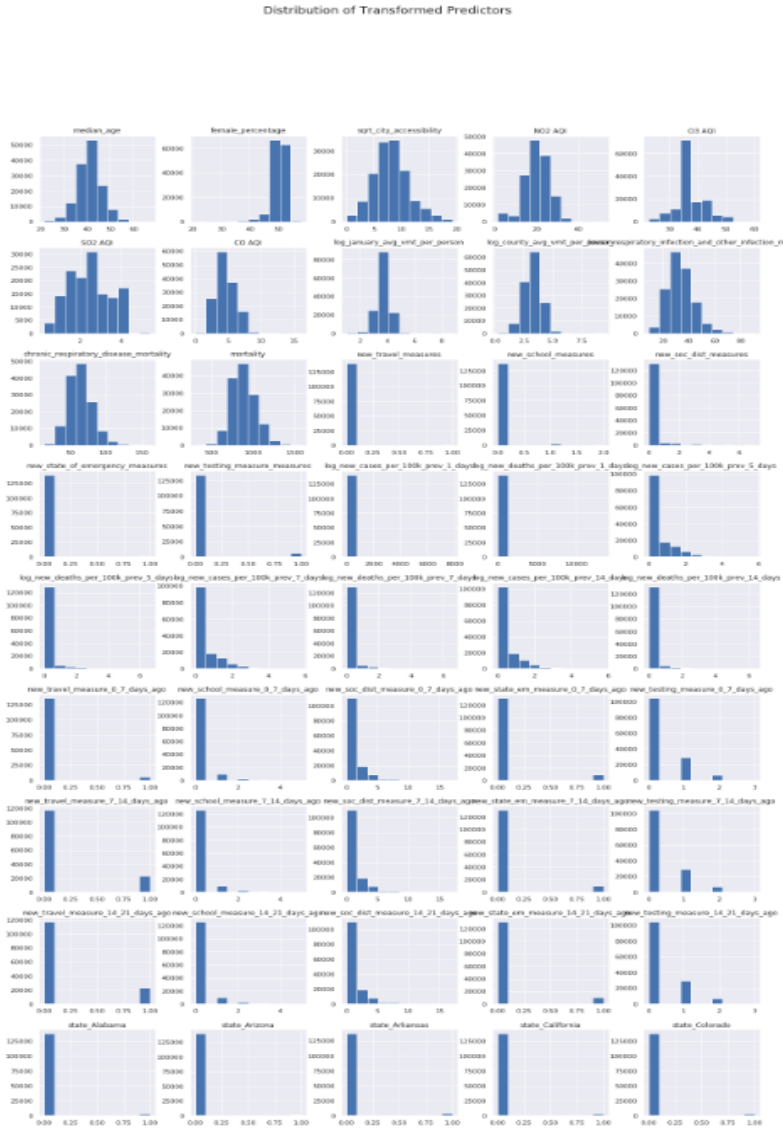
Appendix

A
Correlation Matrix of Features



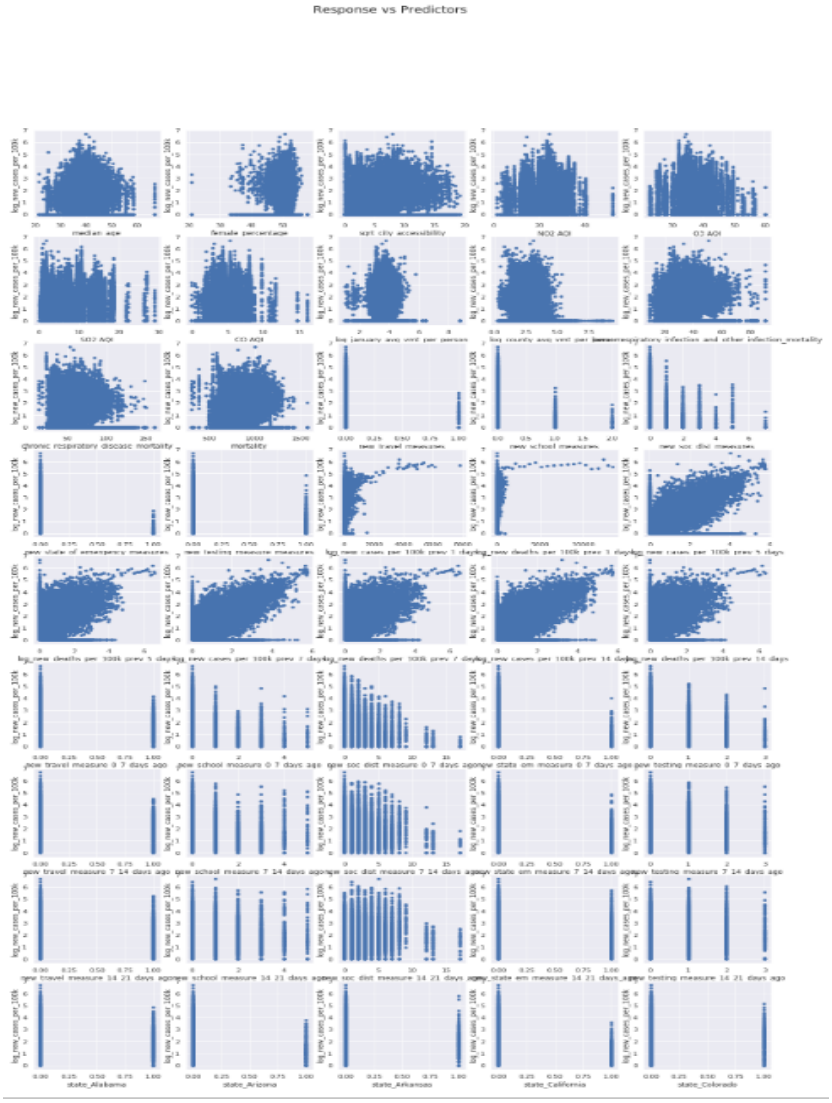
Correlation is quite low with the exception of the rolling window features.

B
Distribution of Transformed Predictors



Following applied transformations, distributions are much closer to normal for the features.

C
Scatterplot of Transformed Response vs Transformed Predictors



The transformations have improved the scatterplots of the data to show more linear relationships.

References

[1] "Ḧow Far Behind The Virus' Is U.S.?: Lag Time In Results Can Hurt Efforts To Control Spread Of COVID-19" *Kaiser Health News*, Kaiser Family Foundation, 2 Apr. 2020, khn.org/morning-breakout/how-far-behind-the-virus-is-u-s-lag-time-in-results-can-hurt-efforts-to-control-spread-of-covid-19/.

[2] Koerth, Maggie, et al."Why It's So Freaking Hard To Make A Good COVID-19 Model" *FiveThirtyEight*, FiveThirtyEight, 31 Mar. 2020, fivethirtyeight.com/features/why-its-so-freaking-hard-to-make-a-good-covid-19-model/.

[3] Laporte, John."Topic: Coronavirus Disease (COVID-19) Pandemic", *Statista*, 2020, www.statista.com/topics/5994/the-coronavirus-disease-covid-19-outbreak/.

[4] Lauer, Stephen A."The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application" *ACP Journals*, ACP, 5 May 2020, annals.org/aim/fullarticle/2762808/incubation-period-coronavirus-disease-2019-covid-19-from-publicly-reported.

[5] "Linear Mixed Effects Models." *Statsmodels*, 2020, www.statsmodels.org/devel/mixed_linear.html.

[6] Oinkina."Understanding LSTM Networks' *Understanding LSTM Networks – Colah's Blog*, Github, 2015, colah.github.io/posts/2015-08-Understanding-LSTMs/.

[7] "Ordinary Least Squares in Python" *DataRobot*, DataRobot, 2018, blog.datarobot.com/ordinary-least-squares-in-python.

[8] Schreiber-Gregory, Deanna. "Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets." *Henry M Jackson Foundation*, 2018, www.lexjansen.com/wuss/2018/131_Final_Paper_PDF.pdf.