Non-Pharmaceutical Response and Prediction Analysis

Raj Kumar Anand, Xin Zeng, Diego Zertuche

March 2021

1 Executive Summary

COVID 19 pandemic unfolded in an unprecedented fashion. Never in our history we had a pandemic that spread across the globe in such a short-time. The spread of the pandemic can be attributed to globalization which has increased the mobility of goods and human resource across the globe. The major challenge in this datathon was to formulate a feasible problem statement which can be answered (in some extent) using the data provided. As a first step, we started with determining the benchmark that can be used to gauge the extent/impact of COVID 19 in any location. This was important because the data had many different variables like 'total cases', 'total deaths' etc which are all correlated and can be potentially used as the target variable. But each of these variables, although correlated among themselves, alludes to different aspect of the pandemic. We resolved this problem by developing an index/benchmark using Principal Component Analysis which tends to capture most of the variance.

As the pandemic progressed, countries started to react to the growing cases and infection rates by deploying non-pharmaceutical response measures while pharmaceutical responses were being developed. Having a good understanding of which responses are the most effective for curving the spread of a disease is crucial for governments as it allows them to make more informed decisions based on the efficacy and cost of a measure. To address this, we analyze how different non-pharmaceutical response measures affect the reproduction rate of the disease in each country and as a whole. We then evaluate which are the most effective measures and which variate most between countries and make recommendations based on these factors. We note that effect on reproduction rate of measures that rely on voluntary basis variate the most between countries, while mandatory measures tend to have a more uniform impact between countries.

To give a more accurate and robust prediction on how the worldwide disease and disease across continents will behave in the future, we not only build Time Series model, Linear Regression model and Gradient Boosting Decision Tree model to compare their performance, but also creatively develop a framework for feature selection. Our framework is robust in worldwide prediction and predictions on Asia and North America. By implementing our framework, we could greatly reduce the number of predictive variables while improving the out-of-sample model performance. With the comparison of different models, we conclude that Linear Regression model with feature selection and without polynomial term consistently performs the best in predicting both the worldwide disease and disease across continents. Because we have small number of data set, general increasing trend of total cases and variables related to vaccinations are all split into testing data set, we observe that complicated machine learning models like GBDT performs worse than Linear Regression model and Time Series model consistently.

With a robust prediction model, we can use the data for current capacity of healthcare services to implement the best Non-Pharmaceutical Response. The implementation of these responses will be a function of the current 'stringency index', 'capacity of hospitals' and 'number of cases predicted in the short-term'.

2 Target Variable Selection

The data set provided us with many different variables that can potentially be used to track COVID. Some of these variables are *total_cases*, *new_cases*, *total_deaths*, *new_deaths*, *new_deaths_smoothed*, *to-*

tal_cases_per_million, *new_cases_per_million*, *new_cases_smoothed_per_million*, *total_deaths_per_million*, *new_deaths_per_million* and *new_deaths_smoothed_per_million*. All these variables are highly correlated among themselves and can be used as a good proxy for ascertaining the state of COVID (with some transformation). This behooves us to select a suitable indicator/index for gauging the COVID, which can be used as a dependent variable in all our analysis. From the preliminary analysis, we notice that all the above variables are highly positively correlated. The figure 1 shows the correlation for the world data.

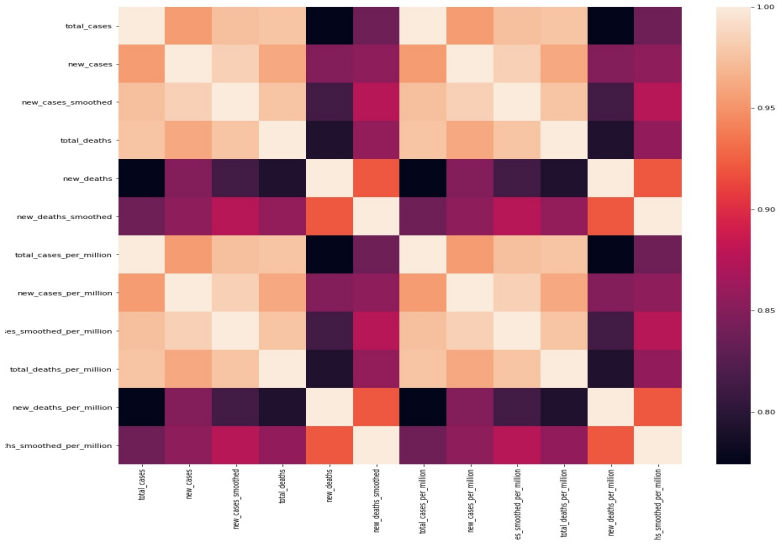

Figure 1: Correlation Heatmap of Potential Dependent Variable for Tracking Covid

Because of the high correlation, we need to create a composite index that can capture the variance of all these variables and can potentially be used as our target (dependent) variable. That's where Principal Component Analysis (PCA) comes in the picture.

2.1 Principal Component Analysis

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

Let $\mathbf{X}$ be the matrix of all the independent variables where each column in a potential target variable and each row corresponds to a date. We then calculate the mean for each column and subtract the respective mean from each variable. This gives us a matrix whose columns has zero mean. We then divide each data with the standard deviation of each columns to center the matrix. We call this centered matrix $\mathbf{Z}$. Covariance matrix of $\mathbf{Z}$ is then calculated by multiplying it with Z^T, which is the transpose of matrix $\mathbf{Z}$.

After obtaining the covariance matrix of the centered matrix $\mathbf{Z}$, we then find its eigen values and eigen vectors. Since the covariance matrix is a semi-definite matrix, the eigen values and eigen vectors can be calculated by eigen value decomposition. The eigen decomposition of Z^TZ is where we decompose Z^TZ into PDP^{-1}, where P is the matrix of eigenvectors and $\mathbf{D}$ is the diagonal matrix with eigenvalues on the diagonal and values of zero everywhere else. The eigenvalues on the diagonal of $\mathbf{D}$ will be associated with the corresponding column in $\mathbf{P}$ — that is, the first element of $\mathbf{D}$ is λ_1 and the corresponding eigenvector is the first column of $\mathbf{P}$. This holds for all elements in D and their corresponding eigenvectors in $\mathbf{P}$. We will always be able to calculate PDP^{-1} in this fashion. Take the

eigenvalues $\lambda_1, \lambda_2, \lambda_n$ and sort them from largest to smallest. In doing so, sort the eigenvectors in **P** accordingly. We call this sorted matrix of eigenvectors **P***. (The columns of **P*** should be the same as the columns of **P**, but perhaps in a different order.) Note that these eigenvectors are independent of one another.

We then calculate the matrix with principal components **Z*** = **ZP***. This new matrix, **Z***, is a standardized version of **X** but now each observation is a combination of the original variables, where the weights are determined by the eigenvector. Since columns of **P*** are independent, the columns for **Z*** are also independent.

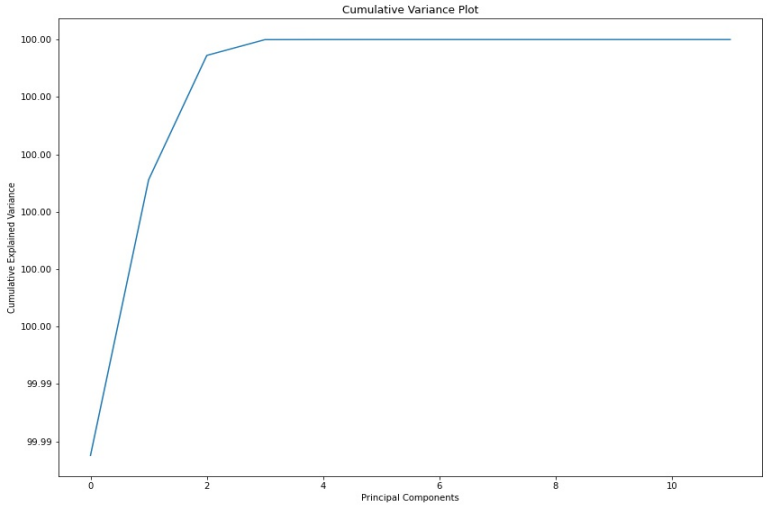

Figure 2: Cumulative Variance Explained

The figure 2 shows the cumulative variance explained by the Principal Components. As evident from the plot, the first principal components explained about 99.99% of the variance.

We then transform our data to calculate the first Principal Component and plot it against 'Total Cases". As seen in the figure 3, total cases is very strongly correlated with our first Principal Component, which explains about 99.99% variance of all the potential target variable. So, using Total Cases as our target variable in further analysis makes sense.

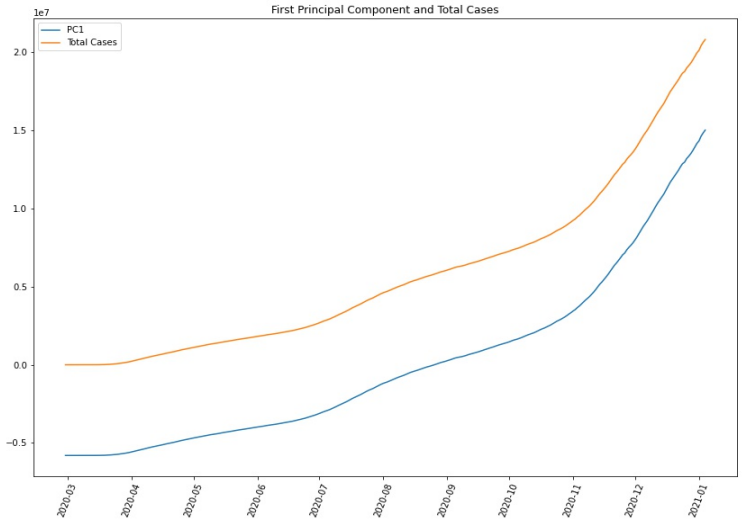

Figure 3: First Principal Component vs Total Cases

3 Non-Pharmaceutical Responses Analysis

As the pandemic progressed, countries started to react to the growing cases and infection rates by deploying a gamma of non-pharmaceutical response measures while vaccines and medical responses were being developed. Understanding which type of non-pharmaceutical response is the most effective in controlling the spread of a disease can be very fruitful for addressing and combating optimally future pandemics that may arise. Having a quantifiable impact of each response and coupling it with the economic, social, etc. cost of said responses will allow governments to make more informed decisions regarding which measure should be taken and when. This is why in this section why try to address the issue of determining which responses that were measured in the *country_response_measures.csv* ECDC (European Centre for Disease Prevention and Control) data set are the most effective for containing the spread of the COVID-19 virus.

3.1 Feature Selection and Multicollinearity

The ECDC data set contains the non-pharmaceutical measures that have been employed and reported by EU countries and the UK and the start and end date of each measure. This data set was transformed from a event based data set to a time dependent data set, where there was a row for each day from January 1, 2020 to January 4, 2021 for each country. All the response measures where converted into binary variables that indicate if the given response measure was in place in that country in the given date. This transformed data was then joined with the OWID (Our World In Data) data set to get the related COVID-19 statistics in the given day and country.

The response variable chosen to perform the analysis was *reproduction_rate*, which is the estimated reproduction rate, or the expected number of cases directly generated by one case, for each day for each country. In this way we can directly relate how a response measure impacts the contagion between individuals.

Table 1: Features selected and description

Variable	Description
StayHomeOrder	Stay-at-home orders for the general population
StayHomeGen	Voluntary stay-at-home recommendations for the general population
PrivateGatheringRestrictions	Restrictions on private gatherings
ClosSec	Closure of secondary schools
MassGatherAll	Interventions are in place to limit mass/public gatherings
ClosPubAny	Closure of public spaces of any kind
MasksVoluntaryAllSpaces	Protective mask use in all public spaces on voluntary basis
MasksVoluntaryClosedSpaces	Protective mask use in closed public spaces/transport on voluntary basis
MasksMandatoryAllSpaces	Protective mask use in all public spaces on mandatory basis
MasksMandatoryClosedSpaces	Protective mask use in closed public spaces/transport on mandatory basis
Teleworking	Teleworking recommendation

The ECDC data set contains 64 unique response measures, as the level of granularity was high in the measures (i.e. there was a response measure of closing nurseries, of closing primary schools, closing secondary schools, etc.) and it was obvious that using all responses would encounter problems with multicollinearity and interpretation. We constructed a correlation matrix to understand the multicollinearity of the responses, and the measures were chosen based on this factor and trying to have at least one of every category. The chosen measures are presented in Table 1. Then the correlation matrix for these elected responses (Fig. 1) was calculated to make sure these were sufficiently uncorrelated. Once having the response variable and predictors, a bayesian random slopes model was chosen to realize the analysis.

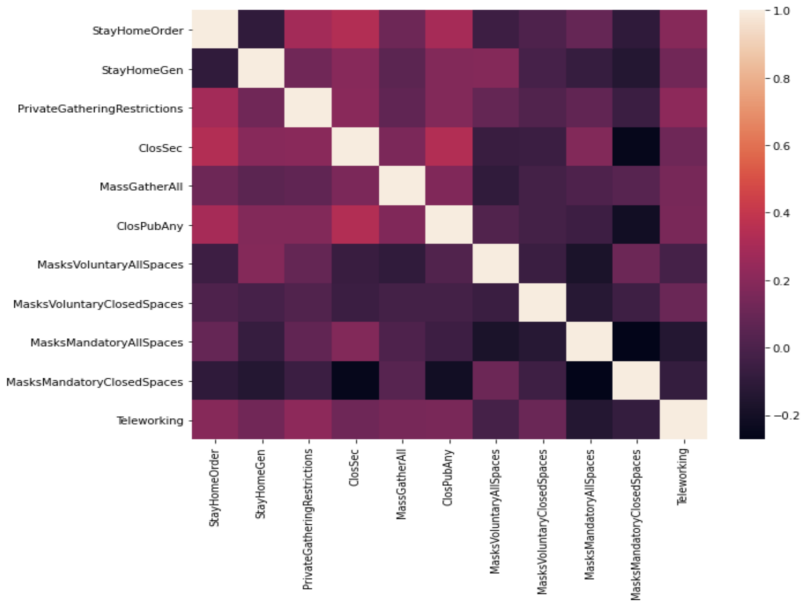

Figure 4: Correlation heatmap of chosen response measures.

3.2 Bayesian Random Slopes Model

A Bayesian Random Slopes model was chosen given the hierarchical nature of the data, as the data is segmented by country. In this way, instead of trying to create a model for each of the countries

that lives independently of each other, we can leverage the information from all the countries in the estimations of the effects of the response measures and have results that generalize better due to the partial pooling that occurs in the model and also have uncertainty measures of the parameters estimated. The prior distribution for this particular model is defined as follows:

$$\mu_r = \beta_{0j} + \beta_{1j} * measure_1 + ... + \beta_{ij} * measure_i$$

$$\frac{1}{\sigma_r^2} \sim Gamma(0.1, 0.1)$$

$$\beta_{0j} \sim \mathcal{N}(\mu_0,\, \sigma_0^2)\ , with\ \ \frac{1}{\sigma_0^2} \sim Gamma(0.1, 0.1)\ and\ \ \mu_0 \sim \mathcal{N}(0,\, 100^2)$$

$$\beta_{1j} \sim N(\mu_1, \sigma_1^2)\ , with\ \ \frac{1}{\sigma_1^2} \sim Gamma(0.1, 0.1)\ and\ \ \mu_1 \sim \mathcal{N}(0,\, 100^2)$$

$$.$$

$$.$$

$$\beta_{ij} \sim N(\mu_i, \sigma_i^2)\ , with\ \ \frac{1}{\sigma_i^2} \sim Gamma(0.1, 0.1)\ and\ \ \mu_i \sim \mathcal{N}(0,\, 100^2)$$

$$R \sim \mathcal{N}(\mu_r, \sigma_r^2)$$

where R is the response variable, *reproduction_rate*, i is the index of the response measure and j is the index of the country. We are making a distribution assumption in the effect of each response measure, where we are assuming that the effect (coefficient) of a response measure is sampled from a same parent Normal distribution for all countries. We are also making a distribution assumption on the hyperparameters of the normal distributions, where the μ is normally distributed and the τ ($\frac{1}{\sigma^2}$) is distributed by a Gamma distribution, this with the purpose of having uncertainty measures in our parameters. In other words, for each response measure, there is a normal distribution that generates the effect of that response for each country, and the hyperparameters of said normal distributions are also being sampled by other distributions. From this model, we can choose the granularity of our analysis, an analysis per country (utilizing the distribution results from the coefficients) or an analysis of the overall effect (utilizing the distribution results of the hyperparameters).

To support our distribution assumptions for the coefficients, we decided to perform multiple country-level OLS regressions utilizing bootstrapped data, with the form

$$R = \beta_0 + \beta_1 * measure_1 + ... + beta_i * measure_i$$

to get a sample distribution for each of the coefficients of the response measures. The distributions for all coefficients were approximately normal (see Figure 2), which supported our distributions assumptions. The prior distributions for the hyperparameters where decided to be non-informative, which is why we gave them $\mu \sim \mathcal{N}(0,\, 100^2)$, as this gives a very flat distribution which gives essentially an equal probability density to all numbers. The same concept applies to the prior distribution of σ^2, where a Gamma distribution was chosen to ensure that the sampled values are always positive (because by definition variance can't be negative).

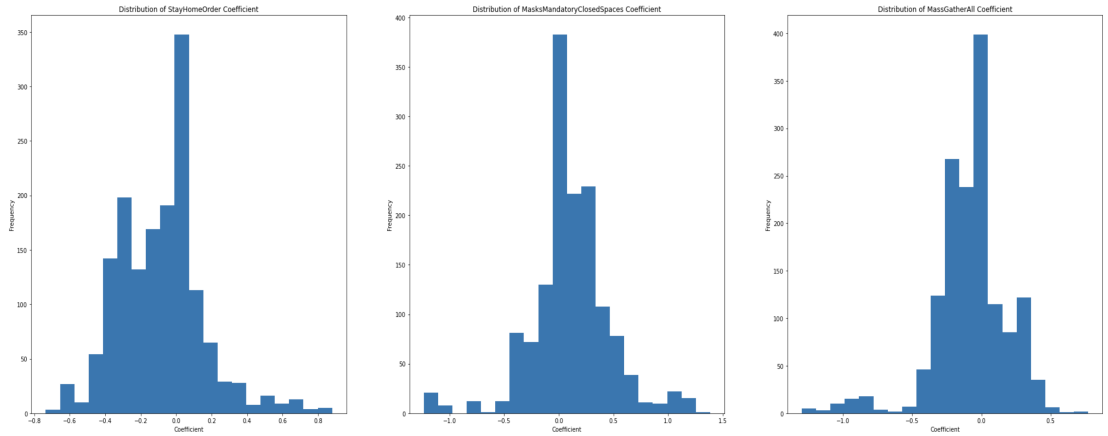

Figure 5: Sample distribution of OLS coefficients for *StayHomeOrder*, *MaskMandatoryClosedSpaces*, *MassGatherAll*

Because calculating the posterior densities of all the parameters of the given model analytically is very complex, we decided to approximate sample values of the posterior densities utilizing Markov Chain Monte Carlo simulations (MCMC)[1]. This method consists on creating a Markov Chain with a stationary distribution that is the same as the posterior distribution of the given parameter, then simulate values from that Markov Chain and consequently obtaining the distribution statistics from those sample values. The Markov Chains were constructed via Random Walk Metropolis (RWM) sampling[2] and utilizing 4 parallel chains.

3.3 Results

3.3.1 Convergence

The MCMC samplers were run with the defined model and the trace plots, which show the history of the value of a parameter across iterations of a chain (Fig. 3) and the $\hat{r}$ statistics, the ratio of between-chain variance and within-chain variance, for the parameters were examined to determine if the samplers reached convergence. The trace plots were stable and the $\hat{r}$ statistics were 1 for all parameters, which clearly denoted that convergence was achieved.

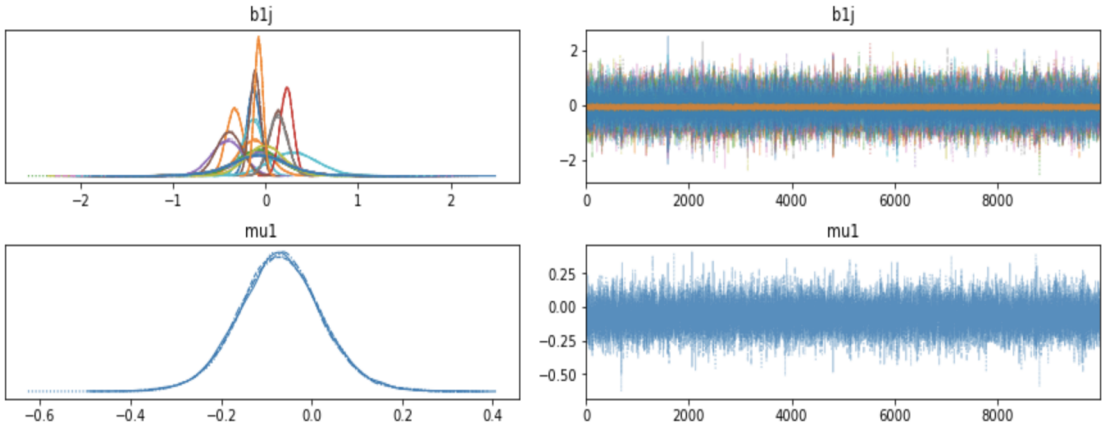

Figure 6: Trace plot and posterior density distribution of the coefficients of *StayHomeOrder* (one for each country) and the μ of the parent normal distribution.

3.3.2 Posterior Statistics

To understand which are the response measures that provide the most benefit in general, we can analyze the posterior distribution statistics of the μs of the parent normal distributions of the coefficients,

as we know that the μ represents the mean in a normal distribution. The posterior mean, standard deviation, and 94% high density interval (HDI) of the μ of the parent distribution for the coefficients of each response measure and the intercept can be found in Table 2.

There are several response measures that we can say with 94% certainty that the average impact in the reproduction rate is negative (therefore reducing transmission). These are *Teleworking*, *Masks-MandatoryClosedSpaces*, *MassGatherAll* and *PrivateGatheringRestrictions*. Out of them all, the one that has the highest impact in reproduction rate seems to be *Teleworking*, as the mean effect it has in reproduction rate is a reduction of -0.282 in the rate, while *MasksMandatoryClosedSpaces* is the second best with a -0.243 reduction. It has to be noted that there are response measures with higher average reduction, which is the case of *MasksVoluntaryAllSpaces* and *MasksVoluntaryClosedSpaces*, but the standard deviation is of the average effect is too high, therefore the HDIs are too wide and we can't conclude that the effects are truly negative. Taking into account the type of measures with this phenomenon, we can hypothesize that this happens because when you have only a recommendation of wearing masks in place, it depends on the culture and other factors of the country if the population wears them, thus this response having a higher variance between countries and its effectiveness being more dependent on country factors.

Table 2: Posterior statistics of μs

Variable	Mean	Standard Deviation	3% HDI	97% HDI
Intercept μ	1.773	0.126	1.535	2.01
StayHomeOrder μ	-0.07	0.093	-0.245	0.106
StayHomeGen μ	0.098	0.160	-0.209	0.396
PrivateGatheringRestrictions μ	-0.188	0.069	-0.317	-0.056
ClosSec μ	-0.078	0.052	-0.177	0.019
MassGatherAll μ	-0.143	0.072	-0.277	-0.006
ClosPubAny μ	0.100	0.099	-0.090	0.282
MasksVoluntaryAllSpaces μ	-0.277	0.216	-0.692	0.126
MasksVoluntaryClosedSpaces μ	-0.300	0.272	-0.809	0.221
MasksMandatoryAllSpaces μ	-0.152	0.088	-0.313	0.019
MasksMandatoryClosedSpaces μ	-0.243	0.128	-0.484	0.000
Teleworking μ	-0.282	0.133	-0.533	-0.033

3.3.3 Model Fit

To test the sanity of our model, we decided to quantify the model fit to the data by taking building out the model with the posterior means of the parameters, and testing the accuracy of the model in out of sample data. The r^2 of the model in test data was of 0.554, which was deemed as a trustworthy result for inferences purposes.

4 Prediction Analysis

In this part, we not only build Time Series model, Regression model and Gradient Boosting Decision Tree model to predict how the disease will behave in the future, but also develop a framework for feature selection. We compare the performance of our model when applying it worldwide and across continents. And we find that Linear Regression model with feature selection and without polynomial term consistently performs the best in predicting both the worldwide disease and disease across continents. Our feature selection framework is robust and valid for different models and areas. For prediction analysis, we only use covid-related statistics from owid-covid-data.

4.1 Worldwide Prediction

4.1.1 Exploratory Data Analysis

We firstly draw a correlation plot for worldwide data. Obviously, there are two groups of variables, one contains those relating to the number of cases and deaths, another contains indicators relating to social development and population distribution. Variables inside each group are highly correlated with each other. It's not reasonable to use all of them in our model. This motivates us to develop a framework for feature selection, since it's a difficult task to choose a subset of features for our data set.

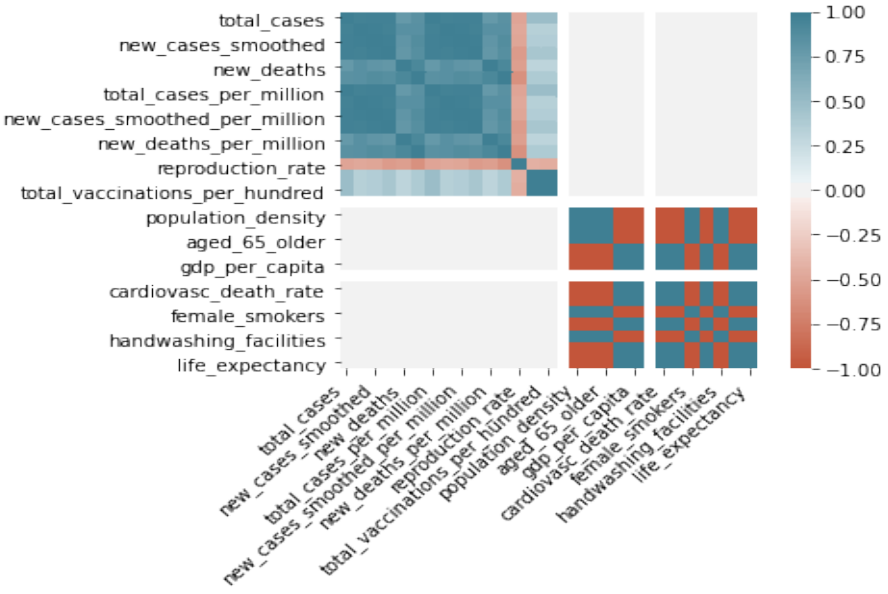

Figure 7: Correlation Heapmap for Data Worldwide.

4.1.2 Data Preprocessing

For worldwide data, we firstly drop the variables that only have missing values. After that, we begin to deal with variables with some missing values, which are *total_vaccinations*, *total_vaccinations_per_hundred*, *reproduction_rate*, *new_cases_smoothed*, *new_cases_smoothed_per_million*, *new_deaths_smoothed_per_million*, *new_deaths_smoothed*. Since worldwide data set here is relatively small, with less than 350 rows, dropping rows due to missing values would have a relatively high influence on our data set. Combined their distributions and the real meaning, we think it's reasonable to fill the missing values with 0.

Besides, to compare the performance of different models, we split our data using uniform way across different models. We first sort our data by date, then split the first 75% of the data as training data, while using the last 25% for testing purposes. And we use RMSE in training and testing data sets for evaluating model performance. MSE is highly biased for higher values while RMSE is better in terms of reflecting performance when dealing with large error values.

4.1.3 Regression Models

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. We firstly fit multiple linear regression for all 29 features. The RMSE on train set is 16407.302, and RMSE on test set is 110225.1956.

Then, we fit a polynomial regression with $degree = 2$ and interaction terms. Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an $n\text{-}th$ degree polynomial in x. An interaction occurs when an independent variable has a different effect on the outcome depending on the values of

another independent variable. And the RMSE on train set is 17569.2143 and RMSE on test set is 1355550.1642.

4.1.4 Gradient Boosting Decision Tree Model

We build a GBDT model using the LightGBM framework[3]. This frame work is high in training speed and efficiency. GBDT is an ensemble tree building method. We start with building a decision tree, and we use the loss of this prediction to update the prediction of the next tree. This means that a subsequent tree learns from the previous mistakes, and the previously inaccurate predictions are given higher weights to be learned.

The set of hyper-parameters we use includes *feature_fraction=0.9*, *bagging_fraction=0.9*, *bagging_freq=5*. This means that for each iteration where a tree is built, LightGBM randomly selects 90% of the features to include in the model. For every 5 iterations, we perform bagging by bootstrapping 10% of the data and leave 90% of the data not resampled. Both of these methods can effectively reduce over-fitting and help the model be more robust. Therefore, we use GBDT as our model for feature selection because it gives a more averaged model that reflects the true effect of the resulting feature space after dropping a variable.

We begin to run a baseline model, which is a single Decision Tree model with $max\_depth = 3$. The RMSE in train set is 15409410.4029, and the RMSE in test set is: 61434506.8469. Then we run GBDT model using all 29 variables as predictors. The RMSE in train set is 344163.6744, and RMSE in test set is 28763945.5363, which shows great improvement than our baseline Decision Tree model.

One advantage of GBDT using the LightGBM framework is that the package easy to implement replicate for future improvement and applications. Also, same as decision tree, it is also easy for GBDT to produce feature importance. Figure 8 is the SHapley Additive exPlanations (SHAP) summary plot for GBDT model trained on our data set. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction[4]. We can find there exists huge difference in feature importance across different features, which motivates us to develop a framework for feature selection.

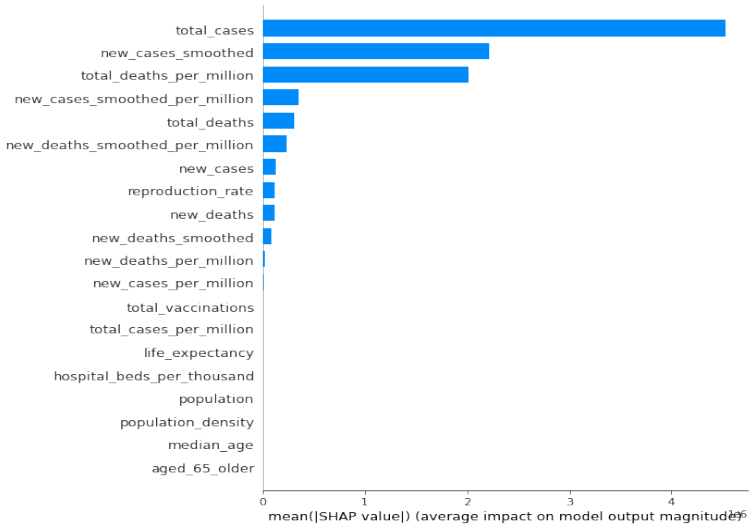

Figure 8: Feature Importance of GBDT Model.

4.1.5 Feature Selection

We aim to choose the smallest set of features which makes our model still robust and accurate. Here, we present a framework for achieving this goal. To figure out the best quantity and combination that we will use, we design a relatively automatic way of selection. First, we split our training data set

with 85% for training and 15% for validation with chronological order. Then, we fit a GBDT model in LightGBM with with 90% feature included each time, for a bagging frequency 5 times and drop the variable with the least SHAP value. Then We iterate the logic until all 29 variables have been dropped, while recording the training and evaluation accuracy. In the meantime, we fit a linear regression and record the training and evaluation accuracy along the way, when the variable is dropped in each step as a benchmark comparison.

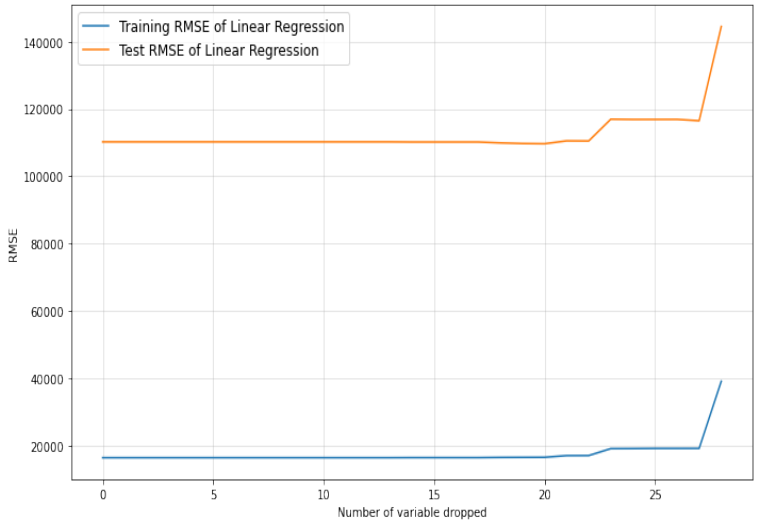

Figure 9: Linear Regression: Relationship Between RMSE and Number of Dropped Variables

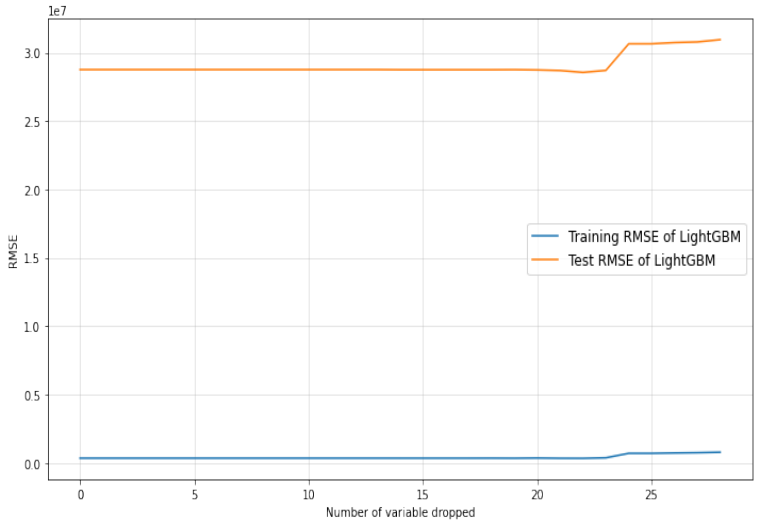

Figure 10: GBDT: Relationship Between RMSE and Number of Dropped Variables

The plots above serve as reference that guides us to the stopping point of dropping. The horizontal label denotes the number of variable dropped. For GBDT model, when the number of variable dropped in less than 24, the training and validation accuracy for stays stable. However, when there are 24 variables dropped, the RMSE increases. Hence, we decide to stop the dropping when 23 variables are dropped, leaving 6 predictors in the reduced model. The 6 predictors are *new_deaths_smoothed_per_million*, *total_deaths*, *new_cases_smoothed_per_million*, *total_deaths_per_million*, *new_cases_smoothed*, *total_cases*. RMSE in train set is 369829.4827, and RMSE in test set is 28698053.4994. Compared with GBDT model with 29 variables, the RMSE in test set has decreased.

For linear regression model, when the number of variable dropped in less than 23, the training and validation accuracy for stays stable. However, when there are 23 variables dropped, the RMSE in-

creases. Hence, we decide to stop the dropping when 22 variables are dropped, leaving 7 predictors in the reduced model. The 7 predictors are *new_cases*, *new_deaths_smoothed_per_million*, *total_deaths*, *new_cases_smoothed_per_million*, *total_deaths_per_million*, *new_cases_smoothed*, *total_cases*. After dropping RMSE on train set is 17040.1007, and RMSE on test set is 110489.1903. We can see we achieve similar performance with only 7 variables as linear regression with 29 variables. We also use the 7 variables for Lasso Regression and Ridge Regression. For Lasso Regression, RMSE on train set is 17045.0874 and RMSE on test set is 110480.7472. For Ridge Regression, RMSE on train set is 17045.1135 and RMSE on test set is 110432.6052.

4.1.6 ARIMA Model

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration. Integration means the use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary. Box and Jenkins (1971)[5] popularised a method which combines both autoregressive (AR) and moving average (MA) models. An ARMA (p,q) model is a combination of AR(p) and MA(q) models and is best used for univariate time series modelling. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA(p,q) model where p is the order of the AR part and q is the order of the MA part.

The parameters of the ARIMA model are defined as follows. p means he number of lag observations included in the model, also called the lag order. d means the number of times that the raw observations are differenced, also called the degree of differencing. q means the size of the moving average window, also called the order of moving average. Here, we set $p = 2$, $d = 1$ and $q = 2$ because this set of parameters would give a stationary estimation. The RMSE on train set is 369699.6555 and RMSE on test set is 10910639.5692.

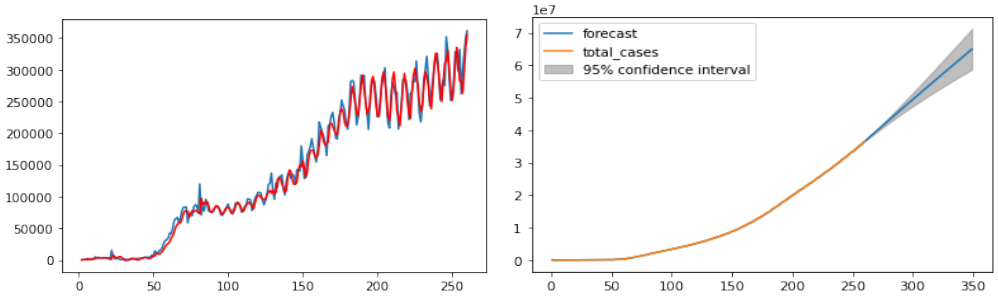

Figure 11: Fitted and Prediction Plot for ARIMA Model

4.1.7 Summary

We summarize our result for different models in the following table. We can observe that Linear Regression without polynomial terms performs the best consistently in training and test data set. GBDT and ARIMA model perform similarly in training data set, but ARIMA gives more accurate estimation in test data set. Besides, we can see that our feature selection framework performs really well. We are able to reduce the variables to 6 or 7 while achieving the same level of accuracy. For GBDT model, by implementing our feature selection, we could even achieve better performance in test data set.

The invalidity of sophisticated machine learning model is due to several reasons. Firstly, since we are doing daily prediction for worldwide data, we only have 349 rows in total. So the amount of data is small. In this case, GBDT model would not be advantageous compared to simple models like Linear Regression and ARIMA model. Besides, *total_vaccinations*, *total_vaccinations_per_hundred* only have 22 non-null values. On the one hand, the general trend of *total_cases* is increasing since the vaccinations are not popular in our case. Besides, since we did worldwide prediction, the variables relating to population and development conditions would not differentiate across different days. On

the other hand, since the 22 non-null values are on the end of the data set. When we split our data set into training and test data set, it would all be split into the test part. Essentially, the vaccination is not considered in our machine learning model. Based on the above analysis, it's reasonable that ARIMA model using a single variable *total_cases* could do better than GBDT model. And Linear Regression model incorporating more variables like those relating to deaths could do better than ARIMA model.

Table 3: Results for Worldwide Prediction

Model Type	Model Details	Training RMSE	Test RMSE
Linear regression	29 Variables, Default	16407.3020	110225.1956
	29 Variables, Polynomial	17569.2143	1355550.1642
	7 Variables, Default	17040.1007	110489.1903
	7 Variables, Lasso	17045.0874	110480.7472
	7 Variables, Ridge	17045.1135	110432.6052
Decision Tree	Max depth = 3	15409410.4029	61434506.8469
GBDT (LightGBM)	29 Variables	344163.6744	28763945.5363
	6 Variables	369829.4827	28698053.4994
ARIMA	order=(2,1,2)	369699.6555	10910639.5692

4.2 Prediction Across Continents

In this part, we follow the similar data preprocessing and modeling steps as we did worldwide previously. As for data preprocessing, we also fill missing values with 0 and split the first 75% of the data as training data, while using the last 25% for testing purpose. As for modeling, we implement Linear Regression and GBDT models for all variables and then do feature selection with our framework. We also include Times Series Model for performance comparison.

4.2.1 Exploratory Data Analysis

By plotting the distribution of total_cases for different continents, we can observe that their distributions are different in across continents, which gives us motivation to build and analyze our model for different continents. In the following parts, we mainly conduct our analysis for Asia and North America because they are representative of two different scale of measurements for dealing with the COVID 19.

Since there are many locations for a single continent, and we hope to make a daily prediction for the continent. We group our data by date and take the summation of each variables across different locations in a single continent. So our response variable for continent analysis is the total values of total_cases for different locations.

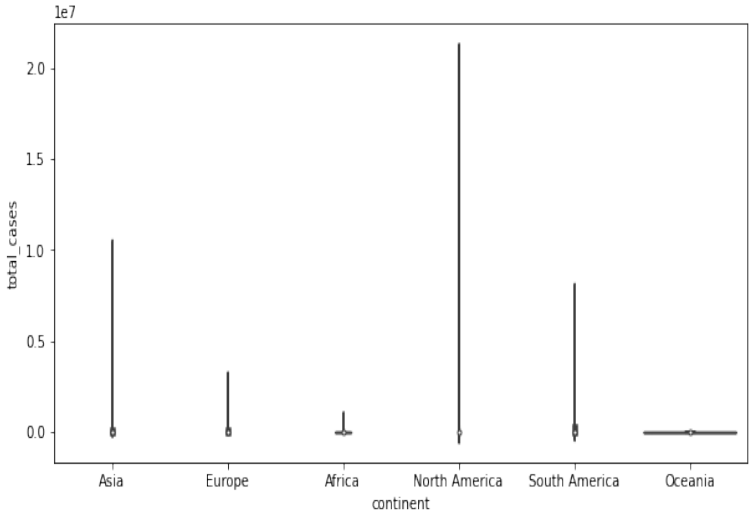

Figure 12: Distribution of Total_Cases Across Continent.

4.2.2 Asia Prediction

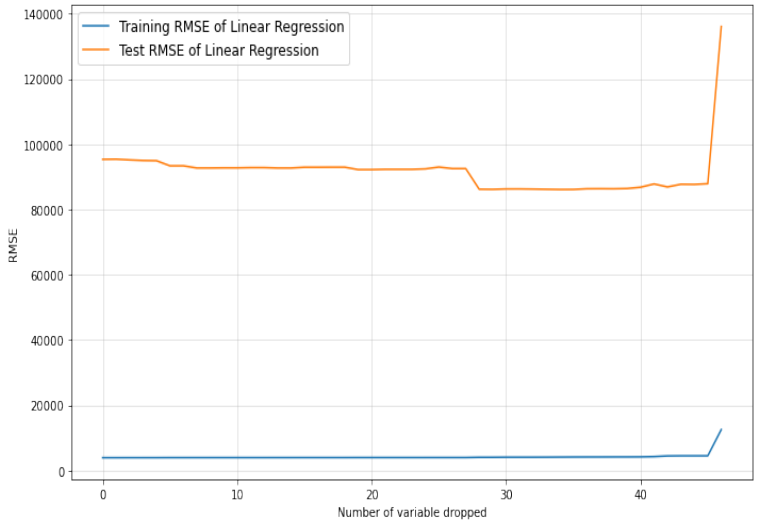

Figure 13: Linear Regression: Relationship Between RMSE and Number of Dropped Variables

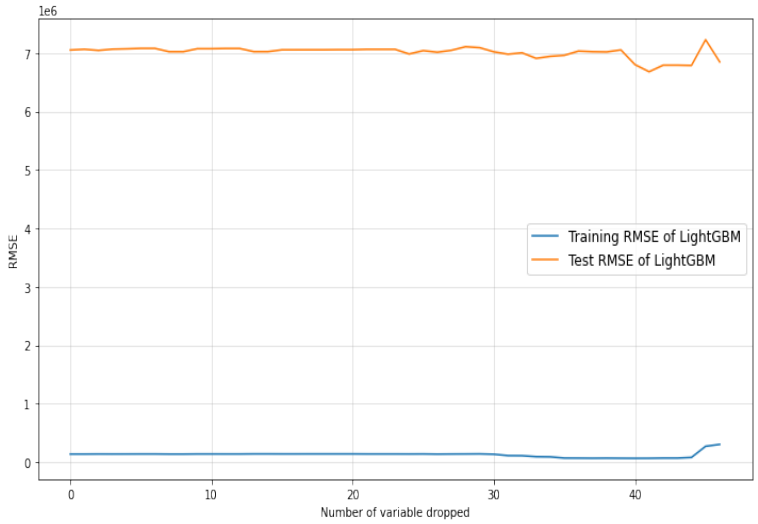

Figure 14: GBDT: Relationship Between RMSE and Number of Dropped Variables.

The plots above serve as reference that guides us to the stopping point of dropping for Asia prediction. The horizontal label denotes the number of variable dropped. For GBDT model, when the number of variable dropped in less than 41, the training and validation RMSE has decreasing trend. When there are 41 variables dropped, the RMSE in validation data set is minimized . Hence, we decide to stop the dropping when 41 variables are dropped, leaving 6 predictors in the reduced model. The 6 predictors are *new_tests*, *total_deaths*, *new_cases_smoothed_per_million*, new_deaths_smoothed_per_million, *new_cases_smoothed*, *total_cases*. After dropping variables, RMSE in train set is 62801.967, and RMSE in test set is 6584904.104. Compared with GBDT model with 47 variables, the RMSE in both training and test set have been reduced.

For linear regression model, when the number of variable dropped in less than 23, the training and validation RMSE stays stable. However, when there are 45 variables dropped, the RMSE begins to increase. Hence, we decide to stop the dropping when 45 variables are dropped, leaving 2 predictors in the reduced model. The 2 predictors are *new_cases_smoothed*, *total_cases*. After dropping the variables, the RMSE on train set is 4485.5974, and RMSE on validation set is 87889.8506. We can see we

achieve better performance in test data set with only 2 variables as linear regression with 47 variables. We also use the 2 variables for Lasso Regression and Ridge Regression. For Lasso Regression, RMSE on train set is 4485.6334 and RMSE on test set is 87943.8071. For Ridge Regression, RMSE on train set is 4485.5974 and RMSE on test set is 87889.8506.

Table 4: Results for Asia Prediction

Model Type	Model Details	Training RMSE	Test RMSE
Linear Regression	47 Variables, Default	3905.3134	95351.905
	47 Variables, Polynomial	45.4392	19709768.6054
	2 Variables, Default	4485.5974	87889.8506
	2 Variables, Lasso	4485.6334	87943.8071
	2 Variables, Ridge	4485.5974	87889.8506
Decision Tree	Max depth = 3	4389814.8233	16485950.1188
GBDT (LightGBM)	47 Variables	137056.4033	7053589.9398
	6 Variables	62801.967	6584904.1040
ARIMA	order=(2,1,2)	102018.8996	391123.8241

We summarize our result for different models for prediction in Asia in the above table. We can observe that Linear Regression without polynomial terms performs the best consistently in training and test data set. GBDT and ARIMA model perform similarly in training data set, but ARIMA gives more accurate estimation in test data set. Besides, we can see that our feature selection framework performs really well. For Linear Regression model, we could reduce RMSE on test data set by reducing the number of predictors to 2. For GBDT model, we could reduce RMSE on test data set by reducing the number of predictors to 6.

4.2.3 North America Prediction

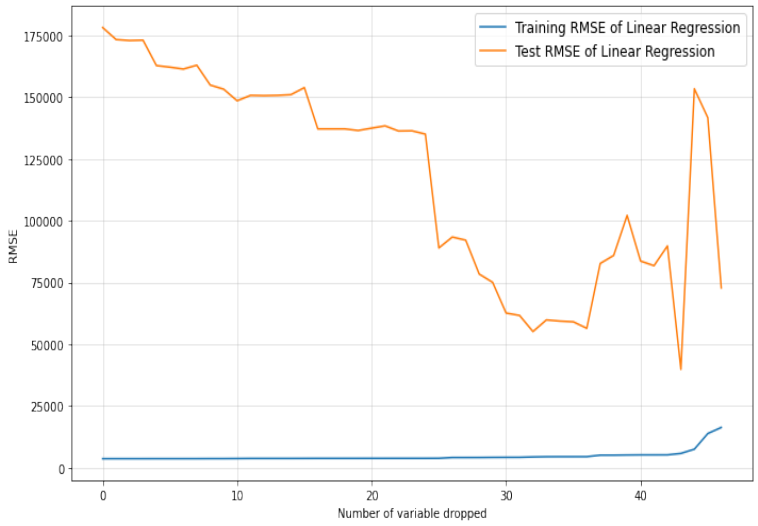

Figure 15: Linear Regression: Relationship Between RMSE and Number of Dropped Variables

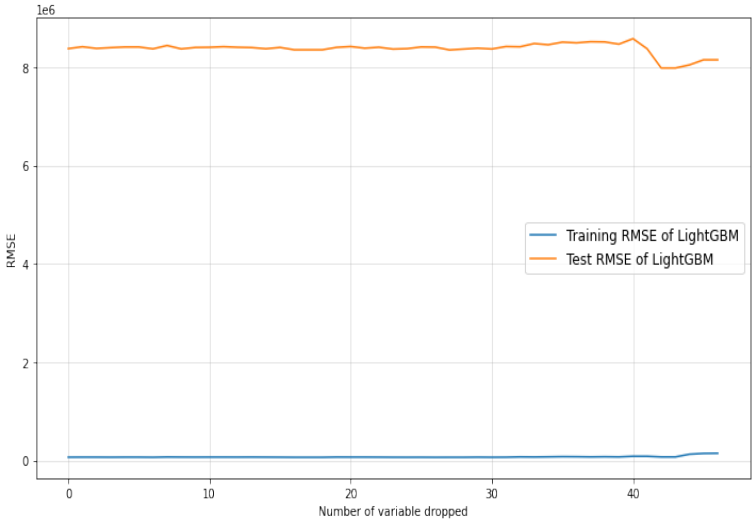

Figure 16: GBDT: Relationship Between RMSE and Number of Dropped Variables.

The plots above serve as reference that guides us to the stopping point of dropping for North America prediction. The horizontal label denotes the number of variable dropped. For GBDT model, when there are 42 variables dropped, the RMSE in validation data set is minimized, and the RMSE in training data set is also relatively small. Hence, we decide to stop the dropping when 42 variables are dropped, leaving 5 predictors in the reduced model. The 5 predictors are *new_cases_smoothed_per_million*, *total_deaths*, *new_cases_per_million*, *new_cases_smoothed*, *total_cases*. RMSE in train set is 70736.7494, and RMSE in test set is 7988608.3625. Compared with GBDT model with 47 variables, the RMSE in test set decreases.

For Linear Regression model, when there are 43 variables dropped, the RMSE in validation set increases. Hence, we decide to stop the dropping when 43 variables are dropped, leaving 4 predictors in the reduced model. The 4 predictors are *total_deaths*, *new_cases_smoothed*, *total_cases*, *new_cases_smoothed_per_million*. After dropping RMSE on train set is 5850.5493, and RMSE on test set is 39852.9804. We also use the 4 variables for Lasso Regression and Ridge Regression. For Lasso Regression, RMSE on train set is 5850.5537 and RMSE on test set is 39986.7189. For Ridge Regression, RMSE on train set is 5850.5493 and RMSE on test set is 39852.9617.

Table 5: Results for North America Prediction

Model Type	Model Details	Training RMSE	Test RMSE
Linear regression	47 Variables, Default	3717.3448	178129.6021
	47 Variables, Polynomial	12.8415	73926369.7338
	5 Variables, Default	5850.5493	39852.9804
	5 Variables, Lasso	5850.5537	39986.7189
	5 Variables, Ridge	5850.5493	39852.9617
Decision Tree	Max depth = 3	4257314.4601	15836110.7213
GBDT (LightGBM)	47 Variables	65383.0333	8376865.5767
	4 Variables	70736.7494	7988608.3625
ARIMA	order=(2,1,2)	54991.0886	4452191.4017

We summarize our result for different models for prediction in North America in the above table. We can observe that Linear Regression without polynomial terms performs the best consistently in training and test data set. And ARIMA gives more accurate estimation in both training and test data set compared to GBDT model. Besides, we can see that our feature selection framework performs really well. For Linear Regression model, we could reduce RMSE on test data set from more than 170000 to about 40000 by reducing the number of predictors to 5. For GBDT model, we could slightly reduce RMSE on test data set by reducing the number of predictors to 6.

4.2.4 Summary

After comparing model performance across continents, we can conclude that Linear Regression model without polynomial terms consistently perform the best. And the performance is robust across continents. Besides, our framework for feature selection is also efficient and valid across continents. We could achieve better prediction results in out-of-sample test data set through our framework, thus providing a powerful tool for dealing with over-fitting and making robust prediction. Besides, our framework tend to be more powerful for prediction in North America, reducing RMSE on test data set from more than 170000 to about 40000 in Linear Regression Model.

5 Conclusion

In this body of work, we examined the different types of non-pharmaceutical response measures and gauged their effectiveness for curving the spread of the COVID-19 virus by reducing the reproduction rate. After performing the bayesian random slopes analysis, we concluded that the most effective response measures were imposing Teleworking restrictions, making masks mandatory on all closed space, banning private gatherings and to limit mass/public gatherings. These were the response measures that we are able to say with 94% confidence that their effect is negative on reproduction rate. The effect on reproduction rate of using masks on closed and all spaces on voluntary basis was found to variate the most between countries, which leads us to believe that non-mandatory measures effect depend more on intra-country factors.

For prediction analysis, we analyze the performance of different models and the validity of our feature selection framework in worldwide predictions and predictions across continents. We conclude that our framework greatly reduce over-fitting and boost out-of-sample model performance. Besides, we find that Linear Regression model without polynomial terms and with feature selection consistently perform the best across different areas.

6 Future Work

The Bayesian Random Slope analysis provided us with the best non-pharmaceutical response that can be implemented to arrest the spread of the virus. While the prediction analysis gave us a robust model to predict the growth of virus spread in future. We can potentially use these models in tandem to time different non-pharmaceutical responses. Based on the severity of the prediction value, which can be gauged from the current capacity of healthcare services, different non-pharmaceutical responses can be implemented. For example, lockdown procedures have been proven successful in mitigating the spread of the viruses in this COVID-19 pandemic, but they also have devastating impact on the economy. A Susceptible-Infectious-Recovered-Deceased model[6] with time dependent infection rate can be implemented to simulate how the infection is spread under lockdown. The economic cost due to the loss of workforce and incurred medical expenses is evaluated with a simple model. This will help us to find the best strategy with minimal cost.

Bibliography

[1] Green P. L. and Worden K. 2015Bayesian and Markov chain Monte Carlo methods for identifying nonlinear systems in the presence of uncertaintyPhil. Trans. R. Soc. A.3732014040520140405 http://doi.org/10.1098/rsta.2014.0405

[2] Gustafson, P. A guided walk Metropolis algorithm. Statistics and Computing 8, 357–364 (1998). https://doi.org/10.1023/A:1008880707168

[3] Lundberg, Scott M, Erion, Gabriel, Chen, Hugh, DeGrave, Alex, Prutkin, Jordan M, Nair, Bala, . . . Lee, Su-In. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding.

[4] Molnar, Christoph. 5.10 SHAP (SHapley Additive ExPlanations) — Interpretable Machine Learning. christophm.github.io, https://christophm.github.io/interpretable-ml-book/shap.html. Accessed 12 Dec. 2020.

[5] Asteriou, Dimitros; Hall, Stephen G. (2011). "ARIMA Models and the Box–Jenkins Methodology". Applied Econometrics (Second ed.). Palgrave MacMillan. pp. 265–286. ISBN 978-0-230-27182-1.

[6] Chao, Sung-Po. Simplified model on the timing of easing the lockdown (September 2020).