MMF1922 DATA SCIENCE PROJECT REPORT

JUNHAO WANG, FEI CHEN, LINGYING SUN

1. Introduction

The Coronavirus disease has had a profound effect on the world since December 2019. The virus is highly contagious and rapidly spreads through close contact via human-to-human transmission. As of October 28, 2020, there have been 43.7 million people infected throughout 219 countries worldwide (World Health Organization). Some individuals have been experiencing mild symptoms such as fever, dry cough and headache whereas others have experienced much more severe symptoms: chest pain, loss of speech, or even death.

The number of infected cases is still growing to this day. To help control the pandemic, each country imposed strict regulations such as travel bans, mandatory masks, social distancing, and the closure of non-essential businesses. Each country's governmental policies have a different effect on the COVID-19 infection rate.

The objective of this project is to build a model to predict future COVID-19 cases and deaths. Five countries' data are analyzed to find the features, which in this study include Australia, Brazil, India, Spain and the United States. These countries were most significantly impacted in terms of number of cases and deaths and so their respective governments have implemented various rules and measures to combat the effects of the virus.

2. Data Set

2.1. **Response Variable.** This model uses current statistics to predict cumulative cases and deaths in two week intervals after a one week period. The first predicted day is a week after today, and it will be added to 14 days. For example, if today is October 1st, based on all the information we have today, the range of our prediction is the cumulative cases and deaths in the period between October 8 to the 22nd. Both future cases and deaths will be studied since they are both practical and meaningful.

The reasoning for predicting the cumulative number of future cases and deaths in a two week period is that the virus will not be tested during its 14-day incubation period. For example, to control the pandemic, the government may impose a rule today. People who have been infected with the virus will likely develop and show symptoms after the 14-day incubation period. Therefore, the influence of the changes today may be in effect 14 days later.

We chose 14 days from today as the center. The two-week total cases and deaths around this day will be investigated as predicted variables.

2.2. **Explanatory Variable.** Our model includes 4 numerical variables and 7 categorical variables which were appended by PySpark.

We used current cumulative cases and deaths as a numerical predictor. Case counts and death counts are only meaningful if we also know how much testing a country administers as the testing number could show the current testing ability of a country which can certainly influence the number of positive cases. Total existing cases were taken into consideration, which is calculated

by subtracting the cumulative number of cured people from the cumulative number of cases. As the number of infected patients increases, the scope of transmission widens, and so the possibility of more infections also increases.
The values of categorical variables listed below are ranged from 0 to 3. The smaller the numbers, the less restrictive the rules are which are detailed below.

Categorical Variables	Measurement	Description
Travel Controls	0	No Restrictions
	1	Close Territorial Boundaries
Restaurant Restriction	0	No Restrictions
	1	Require Takeout Only
Public Events	0	No Restrictions
	1	Recommand Cancelling
	2	Require Cancelling
School Closing	0	No Restrictions
	1	Recommend Closing
	2	Require Closing
	3	Require Closing all levels
Workplace Closing	0	No Restrictions
	1	Recommend Closing
	2	Require Closing for Some Sectors
	3	Require Closing for all-but-essential Workplaces
Stay at Home Requiements	0	No Restrictions
	1	Recommend not Leaving House
	2	Require not Leaving House with Exceptions for "Essential"
	3	Require not Leaving House with Minimum Exceptions (e.g.Allowed to leave onece a week or only one person can leave at a time)

FIGURE 1. Categorical Descriptions

Many countries have enacted travel restrictions in response to the spread of COVID-19. It makes sense that travel control could slow the spread of COVID-19 because the virus could spread via airline routes from severe to mild countries. Secondly, one of the more common ways that COVID-19 is spread is through close contact, like dining in restaurants, because people cannot wear a mask while eating. Restrictions on dine-in may help mitigate the spread. Another a method in which the virus spreads is in public events because events like weddings, funerals and parties usually have larger groups of people confined to a smaller space. Restrictions on public events may reduce the spread and chance of infection. We also see some outbreaks taking place in schools and workplaces. Viruses are spreading between children when interacting with each other at school or between adults when they are speaking with each other. Therefore, school closures and workplace restrictions would have an impact on the number of cases as well. Lastly, staying at home helps control the pandemic because some people do not show symptoms within a span of 14 days. Staying at home reduces the possibility of inadvertently transmitting the disease.

3. Proposed Methodology

3.1. **Data Preprocessing.** Real-world data is often incomplete, inconsistent, and lacking in certain behaviors or trends. It is likely to contain many errors. To tackle this problem, we implemented a data preprocessing technique on our raw datasets which only contains two dependent variables, country, and all numerical independent variables.
Our raw dataset has a lot of missing data. For the total testing, we originally have cumulative weekly testing numbers for some countries since they only record their testing data once a week. We performed linear interpolation using kutools to fill in the missing data in these cumulative

variables because this method works well for a time series with some trend and we converted it into the incremental variable as the input of our dataset. When handling the missing values for other variables that are missing at random, we used pyspark to assign zero to the missing values, with the condition that 0 does not have a meaningful value.
Then, we performed vectorization by rounding on our input dataset. Because of the particularity of our model, all the explanatory values that we give to our model should be in the form of integers.

3.2. **Feature Engineering.** To capture a better understanding of our dataset, we plotted the time series on our dependent variables to visualize their trends.

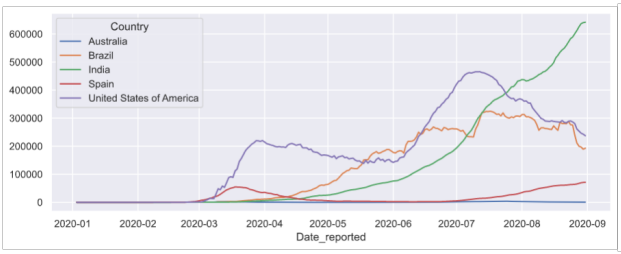

FIGURE 2. Two-week cumulative cases after next week

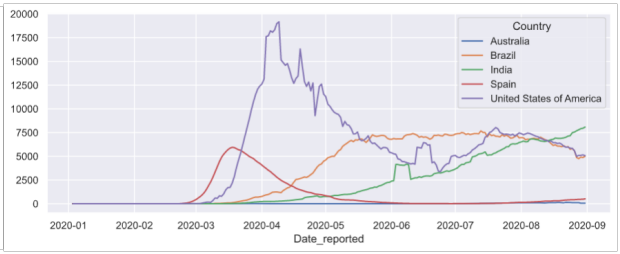

FIGURE 3. Two-week cumulative deaths after next week

From the Figures 1 and 2 which is the time series plot of our two response variables, we could detect, they all approximately increase linearly over time with some fluctuations in between. We suspect these variations may be caused by new restriction rules imposed by the government. So, based on this consideration, we started selecting our features.
We used pyspark to create columns for new features and merge features into our cleaned dataset. Our new dataset contains different categories of classes. We performed one-hot encoding to convert the name of countries into integer representations, which can be handled by our model.

3.3. **Exploratory Data Analysis.** Since the patterns of our two dependent variables are similar, we only investigated the relationship between Two-week cumulative cases after next week and all explanatory variables in this subsection.
The below four Scatter plots could help visualize the relationship between our response variable and numerical independent variables. We are looking for any relationship between the features and cumulative cases. Most countries show positive non-linear relationships except for Spain and the United States. In Figure 4, the more cases have been tested, the more total cases would be confirmed in two weeks. In Figure 5, the graph shows the cumulative cases affect by the total existing cases. We subtracted the recovered cases from total cases, which means the number of people is still sick. We use this feature because if more people are currently sick, the chances of catching Covid-19 in the crowd are relatively large. Figures 6 and 7, both two graphs show the positive relationship between cumulative cases and cumulative deaths except for Spain and the US. There exists a negative relationship at some point. Because we see the variation in the graphs, we believe the variation is coming from other categorical features.

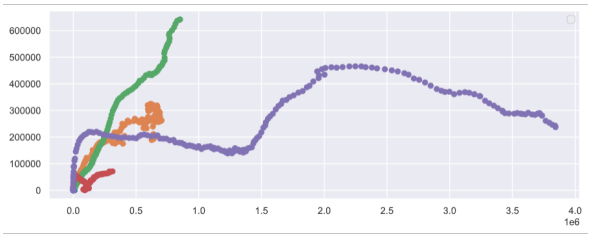

FIGURE 4. Total Existing Cases

FIGURE 5. Total Testing

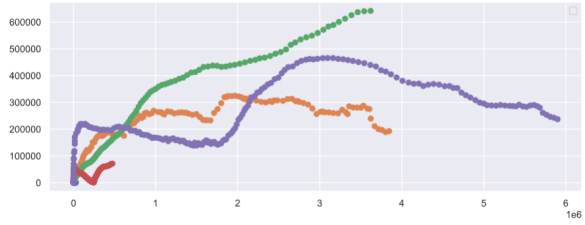

FIGURE 6. Cumulative Cases

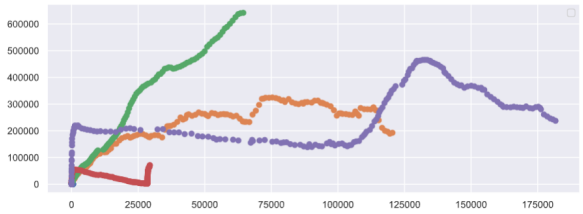

FIGURE 7. Cumulative Deaths

Figure 8 is the scatter plots of our dependent variable and categorical variables, which shows that more restrictions on restaurants could reduce future cases which meets our expectations. For workplace closing and stay at home requirements, different levels of restrictions have different effects. When the rules are strictest, such as closing for all-but-essential workplaces and leaving the house with minimum exceptions, the cases seem to decrease. However, the results in travel controls, school closing, and public events do not make intuitive sense. The figure tells us imposing these rules will lead to more cases. We think maybe the time that the government posed these rules is when the cases are peaking. From these figures, we suspect there are only a few categorical variables that are actually playing an important role in our dataset prediction. However, intuitively imposing restrictions should help control the situation.

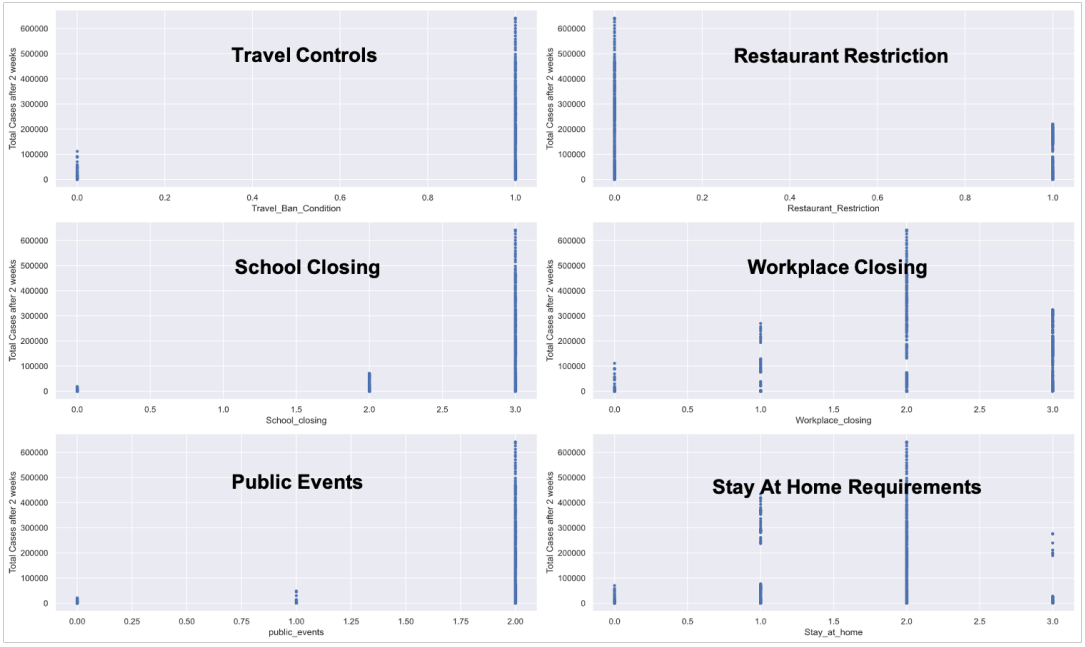

FIGURE 8. Relationship between Two-week cumulative cases after next week and all Categorical Variables

4. Regression Model

4.1. **Lasso Regression Model.** Our dataset only has eleven independent variables to predict our two response variables. Also, from our dataset findings subsection above, we suspected that there are only a few predictors actually influence the two response variables.
As the independent variables we predict are numerical variables, the regression model is chosen as our prediction model. In order to reduce redundant variables and control the overfitting & underfitting, we choose to use L^1 regularization, which is known as Lasso Regression. The term λ is used to penalize the weight, and it is useful to keep the number of parameters of the model large and enforce simpler solutions within the same space of parameters. When $\lambda = 0$, the Lasso regression is equivalent to the simple linear regression. We are aimed to minimize the Residual Sum of Squares (RSS) with the regularization term:

$$\mathscr{L}_{reg} = RSS + \lambda \sum_{i=0}^{k} |\beta_i| = \sum_{i=1}^{n} (Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{i=0}^{k} |\beta_i|.$$

To tune this hyperparameter λ, we use 5-fold cross validation with a set of candidates and find that $\lambda = 5$ is the best hyper-parameter.
After fitting the model, the estimated coefficients for this multivariate regression are:

Explanatory Variables	Symbols	Coefficients
Cumulative cases	$\beta_{1,1}$	0.0907
Cumulative deaths	$\beta_{1,2}$	0.3076
Total Existed	$\beta_{1,3}$	(0.1618)
Total tests	$\beta_{1,4}$	0.3958
School closing	$\beta_{1,5}$	15,132.8433
Workplace closing	$\beta_{1,6}$	(11,706.9183)
Public events	$\beta_{1,7}$	(29,728.6950)
Stay at home	$\beta_{1,8}$	14,139.0791
Travel controls	$\beta_{1,9}$	13,682.4891
Travel Ban Condition	$\beta_{1,10}$	32,724.7292
Restaurant Restriction	$\beta_{1,11}$	(13,019.2250)
Country Australia	$\beta_{1,12}$	(42,496.9996)
Country Brazil	$\beta_{1,13}$	47,987.8357
Country India	$\beta_{1,14}$	(564.2618)
Country Spain	$\beta_{1,15}$	(23,110.5908)
Country US	$\beta_{1,16}$	36,618.1150

FIGURE 9. Coefficients for Y_1 = Two-week cumulative cases after next week

Explanatory Variables	Symbols	Coefficients
Cumulative cases	$\beta_{1,1}$	0.0017
Cumulative deaths	$\beta_{1,2}$	(0.0263)
Total Existed	$\beta_{1,3}$	(0.0011)
Total tests	$\beta_{1,4}$	0.0027
School closing	$\beta_{1,5}$	571.5318
Workplace closing	$\beta_{1,6}$	530.3883
Public events	$\beta_{1,7}$	(801.4202)
Stay at home	$\beta_{1,8}$	176.2036
Travel controls	$\beta_{1,9}$	(18.3980)
Travel Ban Condition	$\beta_{1,10}$	1,222.3489
Restaurant Restriction	$\beta_{1,11}$	728.2072
Country Australia	$\beta_{1,12}$	(948.4004)
Country Brazil	$\beta_{1,13}$	1,868.4246
Country India	$\beta_{1,14}$	(59.2736)
Country Spain	$\beta_{1,15}$	-
Country US	$\beta_{1,16}$	3,223.9833

FIGURE 10. Coefficients for Y_2 = Two-week cumulative deaths after next week

where $\beta_{i,j}$ represents the coefficient of X_i when predicting Y_j

4.2. **R-square Measures.**

R-square of cases prediction is 85.39
R-square of deaths prediction is 60.99

The R-square measures the percentage of variation in the dependent variable explained by the independent variables. The R-square for two-week cumulative cases after next week's prediction is relatively higher than it for two-week cumulative deaths after next week, which means the regression model has a better fit for case prediction than death prediction.

4.3. **Model Predictions.** We have collected the features data of the United States on September 23rd and use the model to predict the cases and deaths. The results are as follow:

The predicted cases are [454682.29128721]
The predicted deaths are [9772.41505048]
True Cases are [677593]
True Death are [9464]

Notice that the predicted cases seems to have a large prediction error while the cumulative deaths are predicted within a reasonable range comparing to the true cases and deaths. The cumulative cases will be influenced by many other factors, and the recent political situations in United States is presumed to be one of the reasons of large cases prediction error. Instead, as a more robust medical care and more government subsidies to the medical system, the situation that a serious infection that lead to the death happens infrequently now.

4.4. **Feature Importance Exploring Using Extra Trees Classifier.** We use an extra-trees classifier to solve regression and problems with both categorical and numerical variables. It helps us to improve the prediction of total cases and death. It works well for over-fitting problems. We also calculated the feature importance to determine how useful the independent variables in our model at predicting future cases and deaths. In figure 9, the feature importance scores show the numerical variables explained the most, and categorical explained the least. We conclude the cumulative cases, total existed cases, cumulative deaths, and total tests are the most useful explanatory variables. Categorical variables didn't explain much but still helping us to predict

the cases and deaths.

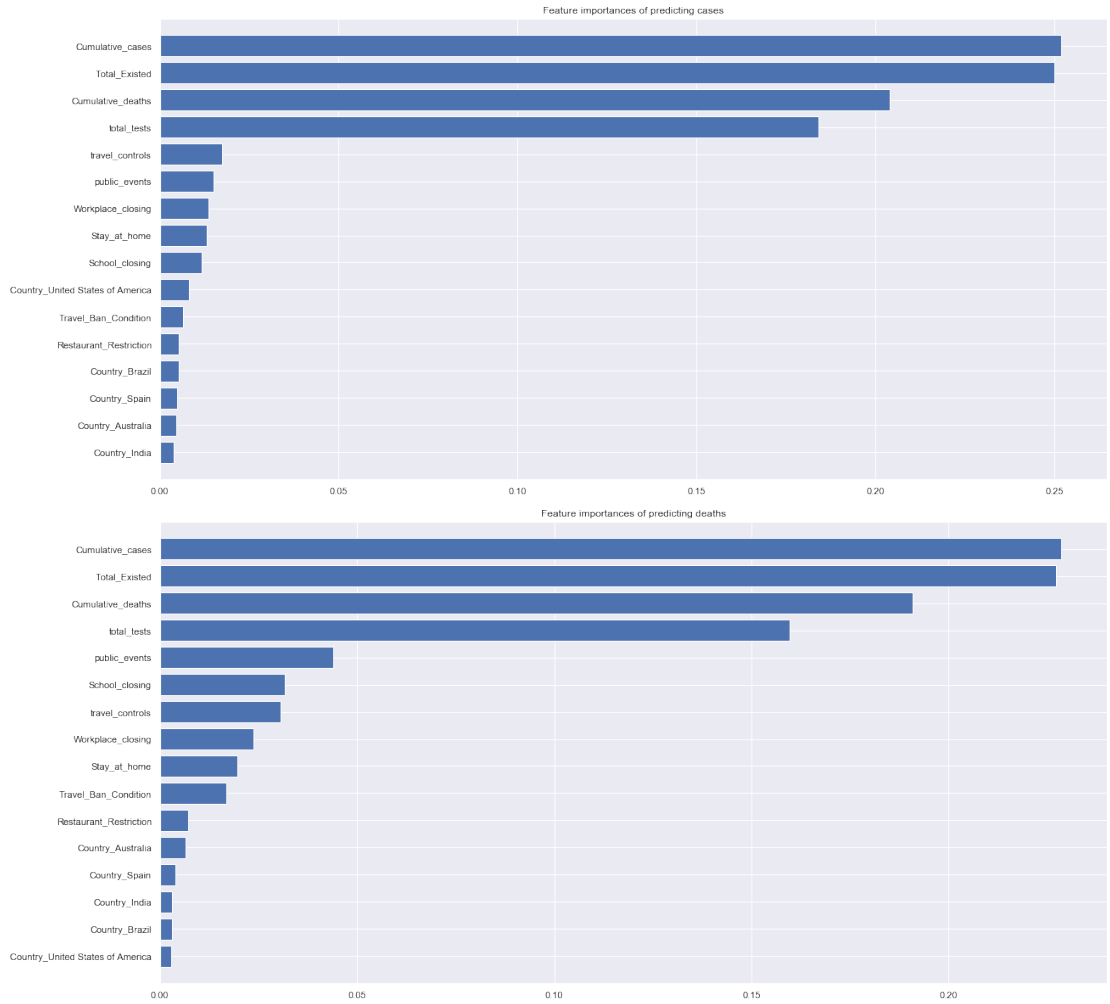

FIGURE 11. Feature Importance Exploring Using Extra Trees Classifier

5. CONCLUSION

Combined with the exploratory analysis and model fitting, we conclude that the future tested cases and deaths will be highly connected to current cases, deaths, total number of tests conducted and many governmental policies. Despite these possible features, people need to pay more attention to the threat of the COVID-19 virus and cultivate stronger self-protection consciousness. Ignoring the virus will be an 'invisible hand' that continuously increases the number of cases and deaths.