Capstone 2 Final Report

Matthew Gudorf

Abstract

The global COVID-19 pandemic has sent shockwaves throughout the entire world and will likely have long-reaching effects on human behavior and culture. There are an incredible number of unknowns and so any investigations into data-driven models are useful. One of the more difficult aspects of this problem is due to the number of unknown factors as well as the time delayed effects. The time delay is due to both incubation time as well as the existence of asymptomatic carriers. Human behavior also contributes to the time delayed effect because people do not instantaneously change their routines. To help governments as well as citizens, I prepare a collection of models which forecast the future number of new cases. As this is mainly an educational project, I only focus on forecasting a single day into the future due to the rate at which these models degrade. To arrive at a quality model, I first examine, clean and wrangle the datasets, explore them, and then test the different model types before creating the final deliverable model.

Description of the data

In order to create a good predictive model I utilize a number of different datasets, choosing and creating numerical features that are believed to contain the most useful subset of information. The original datasets contain information on case numbers, test numbers, government responses, and many other relevant quantities. While there is considerable overlap between these datasets, there is inconsistent reporting and hence not all data are of the same quality. This is examined in more detail in the data exploration notebook; the discussion is too specific to serve any purpose here. These datasets are dynamic; changing everyday to include the most recent information. Originally I was updating these datasets while producing this project but for now I work with fixed data which only goes until June 30th. More specifically I use four different datasets which contain global data, that is, data on a large number of countries around the world. The datasets I use are: the John's Hopkins CSSE (JHU) dataset on cases, recovered, deaths, etc. The "Our World in Data" (OWID) contains case information as well as many time independent quantities such as population, rate of smokers, percentage of population above 65 years of age, etc.. The Oxford COVID-19 Government Response Tracker (OxCGRT) dataset contains the responses of various governments in regards to quarantine measures as well as governmental aid for its citizens. The most important feature is their "stringency index" which scores a government's response using a number from 0-100. The fourth and final data set comes from the FIND test tracker dataset which tracks testing around the world (also contains case

information but this is copied from JHU, per their description). Because I am aggregating multiple datasets, it is important that they are consistent with each other. Towards this end, I apply a number of different wrangling and cleaning steps to produce a dataset which can be explored and used to train my models. I find it the most convenient to use the python package Pandas to perform these tasks.

Data Cleaning and Wrangling

The first of the cleaning steps is to rename the columns and location names such that they are consistent between DataFrames. This is done via a combination of Pandas operations, regular expressions, and custom functions. The datasets all contain different set of reporting days as well as countries. To reconcile this, I elect to take the subset of dates and countries resulting from the intersection of the datasets. The reason for taking the intersection and not the union is to have the fewest number of missing values as possible because the models already degrade quickly as a function of time; any unnecessary sources of error would only compound this issue. Other manipulations that I apply are only to make the data more sensible, such as correcting time series which incorrectly decrease. I know they incorrectly decrease because by definition they are cumulative variables, such as the number of total cases. I do this by repeatedly scanning the time series and propagating the most recent (larger) value whenever an incorrect decrease occurs. Another data quality error is the existence of reporting errors or missing values. This can cause a high amount of irregularity in the time series which seem neither sensible nor accurate. To account for this, I also replace any errant zeros with the most recent value. For instance, if a country is performing 200000 tests a day, it seems incorrect to have a single day where 0 tests are reported. This manipulation might draw criticism, but it does not affect very many values and so I do not believe it has a large effect in hindsight. Perhaps the most important wrangling step I perform is to engineer a number of moving averages of the time dependent features. These features are useful in conveying time dependence to the regression models. I elect to use moving averages of widths 3, 7, 14 days, to capture seasonality and the time delayed nature of the spread of the infection.

Exploration

The COVID-19 pandemic has overshadowed the world since mid to late March 2020, when most countries had at least one recorded case. While the age of the pandemic is approaching four months old for most countries, globally the COVID-19 pandemic seems to be worsening. This is partly due to what I refer to as “quarantine fatigue”. This is a name I give to social unrest which prompts government officials to relax social distancing restriction. To get an idea for how badly the

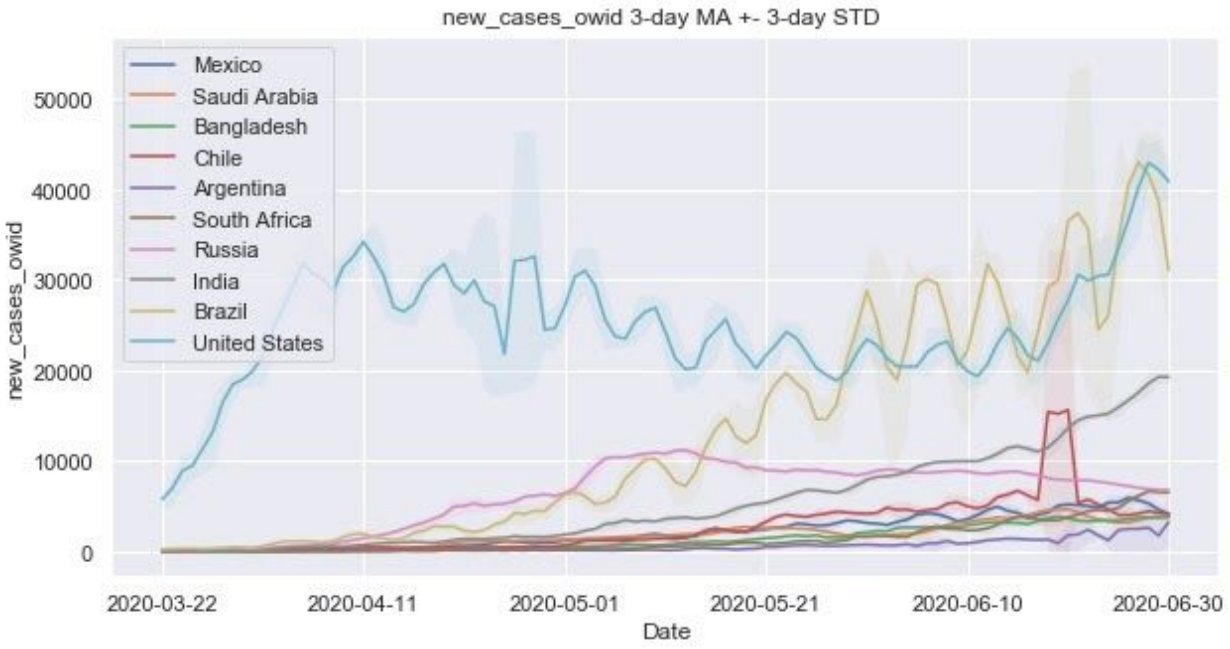

Figure 1: The 3-day moving averages of the 10 countries with highest number of new cases,

current state of the world pandemic is, Figure 1 displays the ten countries which had the highest number of new cases and new deaths (3-day moving average) as of June 30th, 2020. I investigate the first of these qualities by comparing the average reaction time with the current (June 30th) death rate. For the strength of the government's reaction, I investigate whether or not the number of new cases is related to the OxCGRT's "stringency index". Lastly, the different types of responses (school closings, public transport closings, etc.) are investigated by using two dimensional ANOVA analysis to look for statistical relationships between different countries and the types of reactions, using the change in new cases per million people before and after each mandate. A criticism of this usage of Anova is that the different factors in this case are not truly independent, but it is only supposed to serve as a crude measure of the effectiveness of different measures.

I define the "average reaction time" as the average date of the implementation of all government mandates. This can (and is) only computed for countries which indeed enacted all different quarantine measures. The death rate is computed by the number of deaths per capita divided by the number of new cases per capita. Note that positive average reaction time indicates that the country reacted after its first case.

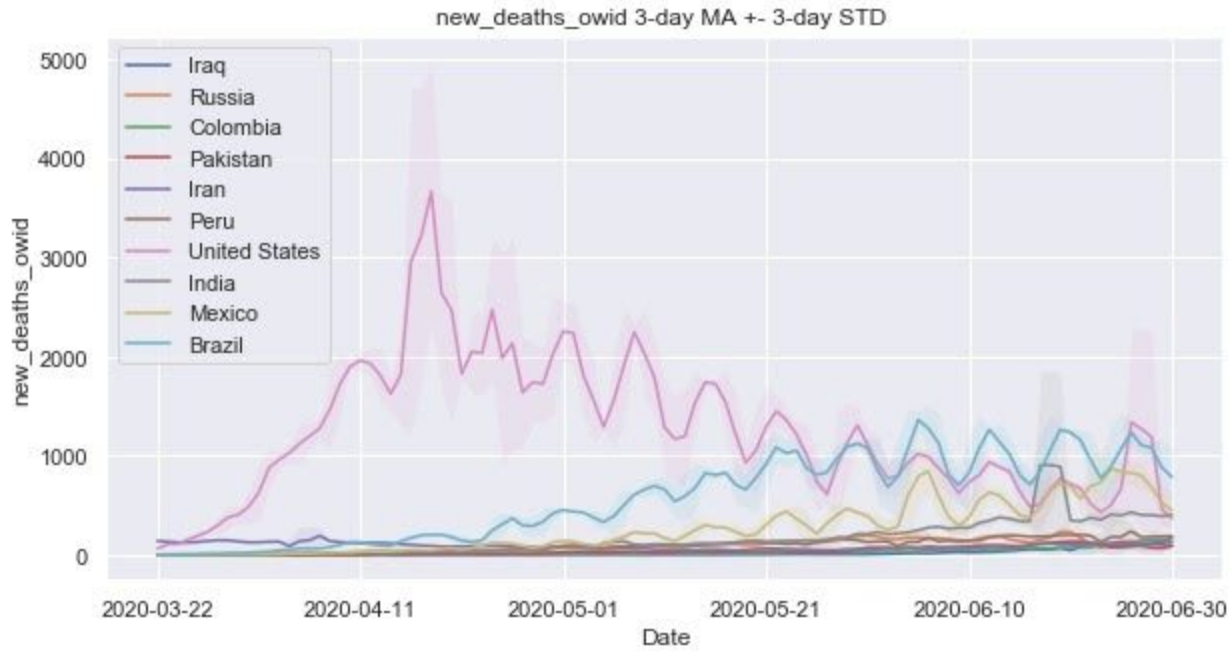

Figure 2 : The number of new deaths time series of the countries from figure 1.

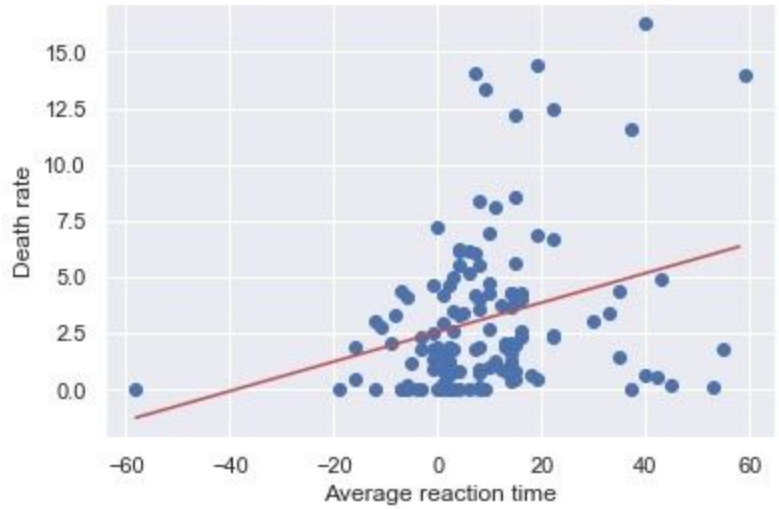

Figure 3 : The current death rate plotted against the “average reaction time” of each government. Average reaction time is calculated by taking the average of the dates of each government mandate.

This provides an intuitive result which can be summarized by the following: countries that are slow to respond to the pandemic suffer more as a consequence. This leads us to my next analysis, given that government intervention is necessary, how strong does said intervention need to be? To explore this, I will first compare two countries which have drastically different outcomes, before looking at the set of all countries.

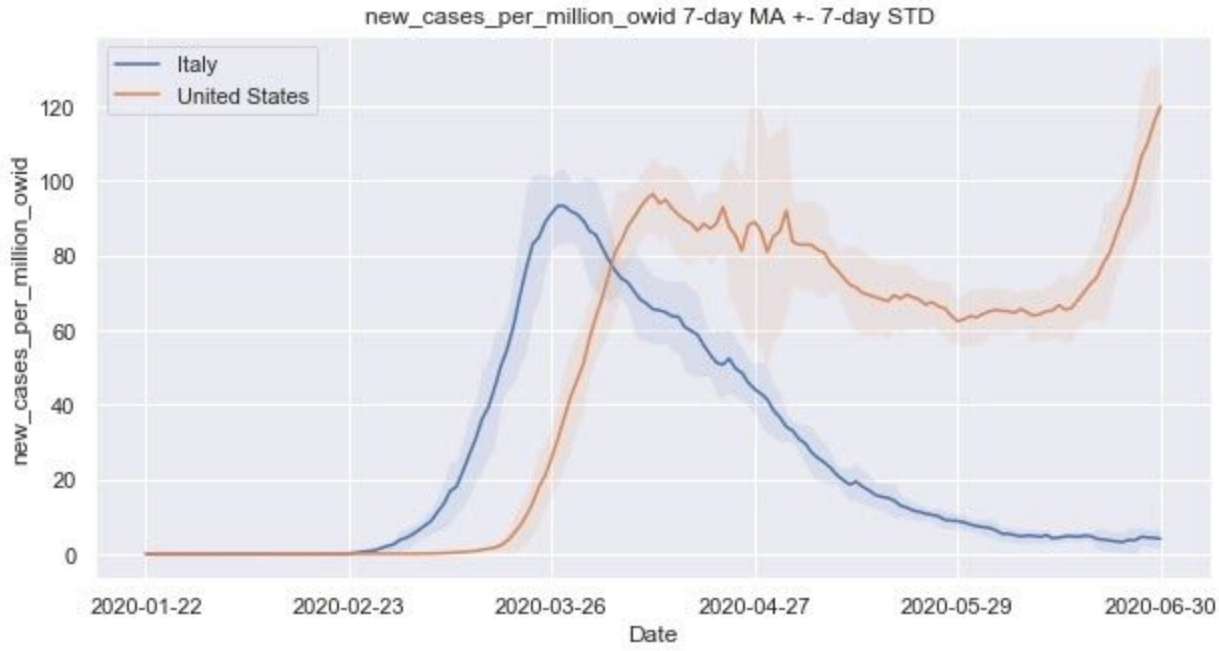

Figure 4: The comparison of the time series for the new cases per million people

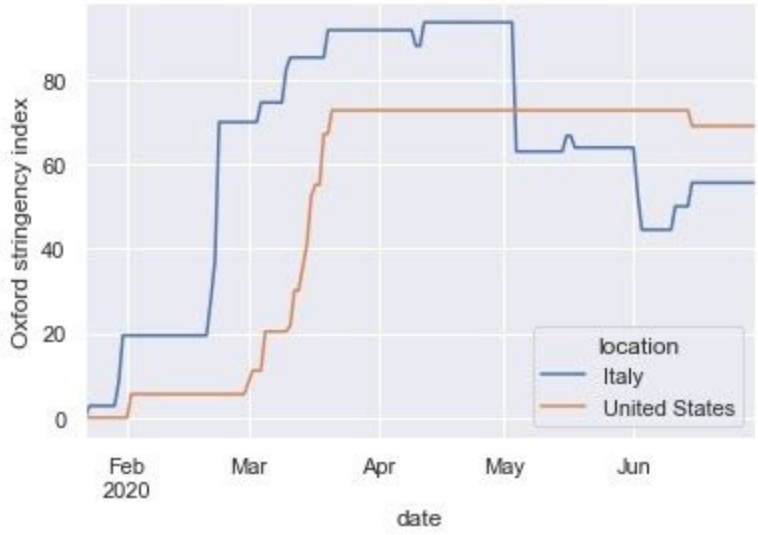

Figure 5: The time series of the OxCGRT "stringency index" for Italy and the United States.

Although Italy had a similar number of cases per million as the United States (and earlier on, when less information on the disease was out), their reaction and subsequent actions seem to have (at least temporarily) quelled the pandemic. As can be seen in Figure 4, the United States likely did not control the pandemic enough, leading to a much smaller decrease followed by a dramatic increase in the number of cases per million. To analyze the differences between each country's reaction, I utilize the OxCGRT dataset's "stringency index", which is a quantification for how each government responded to the pandemic, in terms of quarantine measures. Using Figure 5 to compare the United States and Italy, we see that the Italian government was much more strict when their cases per million were peaking in late March. Once the pandemic was under control, however, this allowed for them to loosen restrictions sooner. This of course is a biased sample of two countries, taken to be a representation of the best and worst case scenarios. I will now attempt to extend this analysis to all countries.

The effect of government actions

The differences between the behavior of the pandemic begs the question, how do government actions affect the spread of COVID19? To investigate this effect, I look to the relationship between the stringency index and the number of new cases (per million). I'm going to utilize 30-day averages in an attempt to only capture the longest time scale and not the seasonality of the pandemic. Specifically I'm going to look at the relationship between the average (and max) stringency index and percentage increase (or decrease) between the average number of new cases two months ago and the previous month. Before jumping right into the problem, I first wanted to get a general idea as to whether this is a plausible hypothesis or not. The easiest way of doing so is to simply plot the global stringency average versus the average number of new cases per million people. As can be seen in Figure 6, they seem to have some inverse proportionality, at least from May forward. Assuming this serves as evidence that my main hypothesis is at least plausible, I begin the full analysis. I engineer three quantities to investigate the effect of stringency. The average number of new cases per million people in the past 30 days, the same quantity between 60 and 30 days ago, and the average stringency between 60 and 30 days ago (used the maximum as well, see COVID19_eda.ipynb). The main quantity I engineer is what I refer to as the "growth factor", the ratio of the average new cases from the two different time periods. A "growth factor" between 0 and 1 indicates a decrease in the number of new cases, and vice versa for values greater than 1. For numerical reasons, I use the logarithm of the growth factors to deal with how tightly clustered the values are in order to be more suited for linear regression. These growth factor values are being compared to the average stringency from 60 to 30 days ago. By using this time period for the average stringency, I am assuming that there is a very significant delay with respect to changes in stringency.

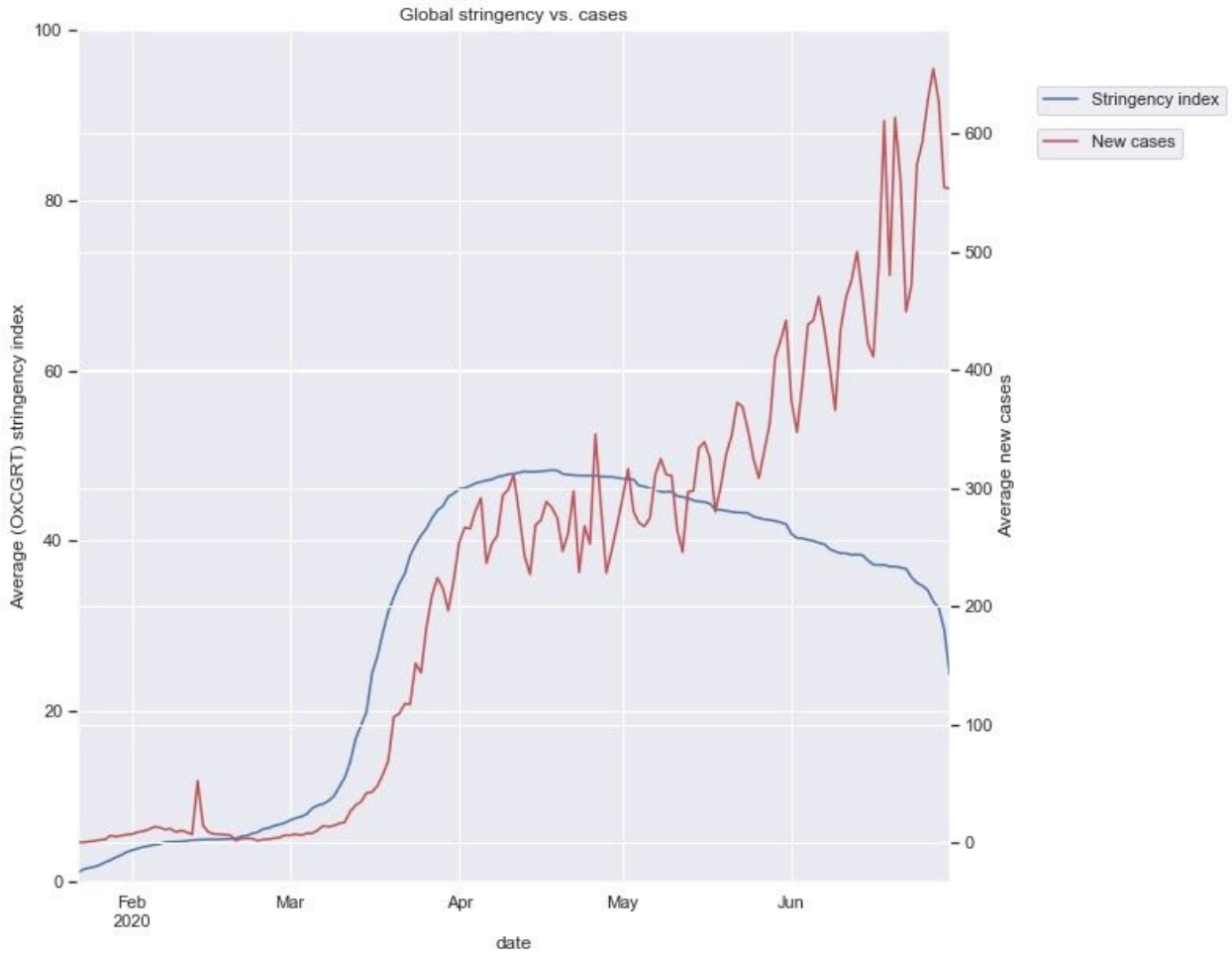

Figure 6 : The OxCGRT government response index vs. the number of new cases per million people, global averages.

The result of my analysis shows that the logarithm of the growth factors does in fact share a relationship with the average stringency index, using a 95% confidence interval. On the contrary, the maximum does not seem to be related in a significant manner, meaning that the old adage “slow and steady wins the race” seems to prove true in this instance.

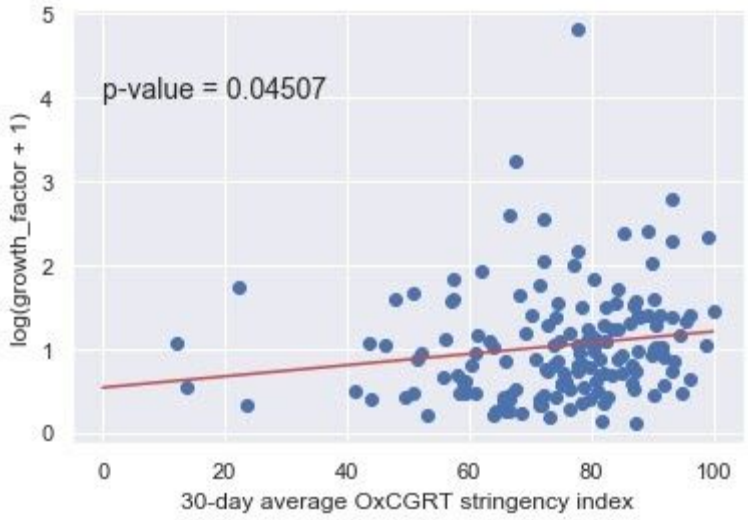

Figure 7 : The regression analysis of average stringency vs. log (growth factors + 1).

The stringency index is an aggregation of different quarantine and social distancing measures. The previous result then makes it obvious to investigate which measures are the most effective. Towards this goal, I use two-way ANOVA analysis, wherein the two different dimensions are the mandate types (closing schools, restricting gatherings, etc.) and the countries. The numerical quantity that serves as the "observation" variable will again be the "growth factors" described previously. In this case, the growth factors are not with respect to the time periods previously used (i.e. 60 to 30 days ago and 30 days ago until the present) but rather they will be computed before and after the implementation of each government mandate. There are a number of countries whose ratios are not finite or defined, therefore these countries are dropped from the ANOVA analysis. Using the statsmodels python package API, I perform type 2 ANOVA analysis on these factors using the ratio values. The result of the analysis is that both the type of quarantine measure as well as the country have an effect on whether the number of new cases drops or not. The ANOVA results are described in the following table, which displays the degrees of freedom, F statistic, p-values and residual values.

	sum_squares	df	F	PR(>F)
Mandate type	0.551935	7	3.008117	0.004305
Country	2.864245	59	1.852094	0.000312
Residual	10.825433	413	NaN	NaN

Table 1: 2-way ANOVA analysis of government mandates

To see whether or not the requirements for the analysis are sensible, I check the distributions of the residuals as well as the variances for the mandate types.

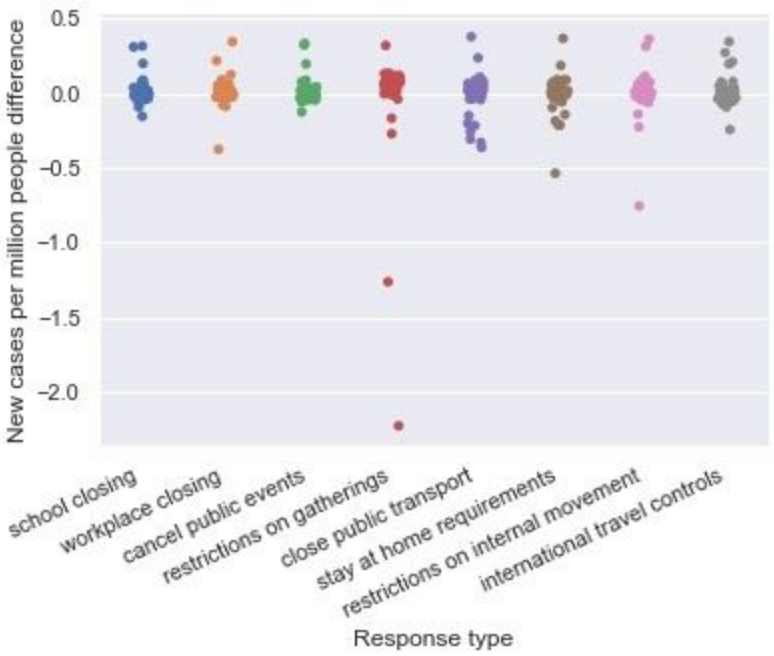

Figure 8: Residual distribution with respect to different factors.

This ANOVA analysis leads me to believe that one of the OxCGRT dataset's index variables (i.e. the government response index or stringency index) should be included in the modeling process. This is important because as it turns out, there are many features which seem to only worsen the performance of the predictive models. I only have hypotheses for why this occurs, but going forward the only features which shall be used (including moving averages for

the ridger regression model) are the number of new cases per million people, its logarithm, and the government response index. The reason why I reduce the number of features so drastically is because there will be at maximum a number of samples on the order of 20000.

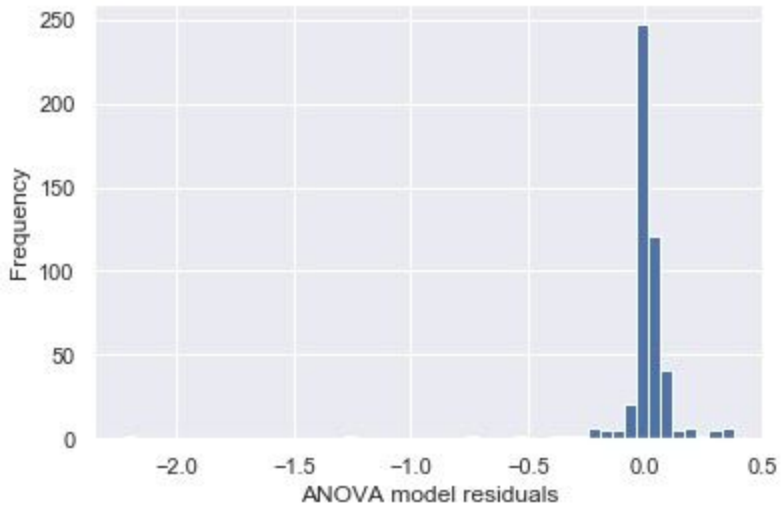

Figure 9: Overall residual distribution

Predictive models for number of new cases per million people

The primary goal of this project is to produce an accurate predictive model for the number of new cases per million people. Naturally, in the context of future prediction, a forecast interval needs to be decided upon. I will focus on predicting a single day into the future. I will utilize three types of models: a neural network with two fully connected layers, a neural network with two convolutional layers and two fully connected layers and a simple ridge regression model. Before attacking the full problem, I developed intuition via a prototyping phase which instantiated the different models, using only a very small subset of the data. The prototyping was completed in the notebook, COVID19_model_prototypes.ipynb. The numerical experiments in this prototyping stage explored the effectiveness of scaling the variables, the utility of differing

data formats, but notably missing was the tuning of hyper parameters. The tuning of parameters was instead saved for the notebooks dedicated to the full version of each model. neural networks so that different combinations could be compared. The data is formatted into different time windows, hereafter referred to as “frames”. The idea is to combine multiple days worth of data in order to make predictions. An alternative to my choice of single-country multiple days format could have also been multiple countries single day or multiple countries, multiple days.

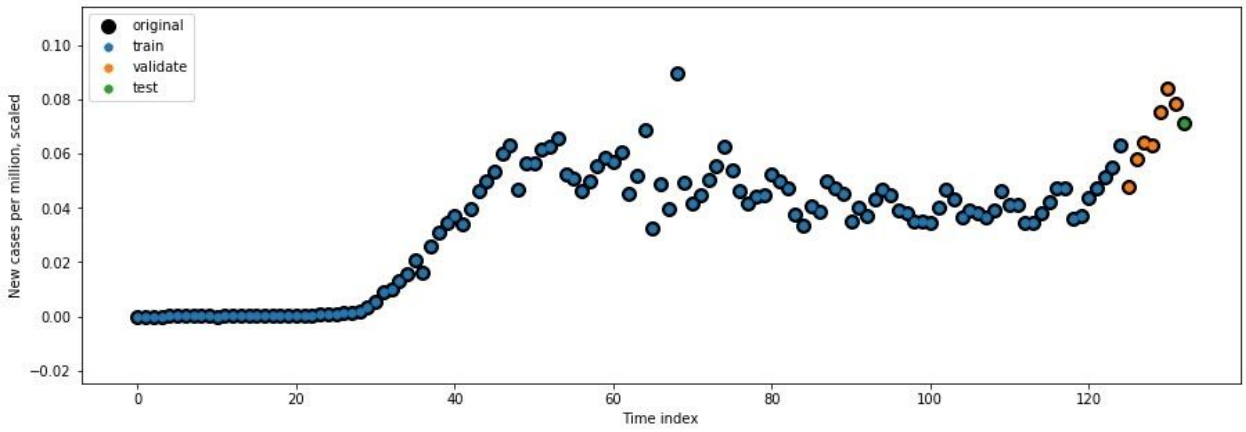

Figure 10 : The split of the different “time frames” into training, validation and testing sets. The dots each represent the most recent date or leading edge of each time frame. In other words, the testing set (green) does not have a single date instead, it contains a single frame of values.

The data is split into time frames, which is shown visually in Figure 10. The color coding represents the splitting of the time frames, where each point in the plot represents the right edge of the time frame. In other words, there is not a single point in the testing set, but a single *frame* whose leading (right) edge is the point colored green. The discussion of the specifics of which features are used and the specifics of the time frames is kept to the separate model notebooks.

I present my analysis visually, in terms of the final performance (predictions) of the three models on the testing set. These predictions are compared with the predictions of a naive baseline, i.e. using the previous day’s values. Each model has a set of four plots. The two plots in the first row are scatter plots of the predicted values vs the actual values. The two plots in the

second row are a plot of the residuals (predicted values - actual values) vs. the actual value. The first column always corresponds to the same naive baseline (it's repeated for each model, for easy comparison). The second (right) column corresponds to either the fully connected neural network ("NN"), the convolutional neural network ("CNN") or the ridge regression model. I used the mean squared error as the loss function for all three models, but also include the explained variance correlation score as well. These scores are included in the top left of each scatter plot of predicted vs. actual values (first row). Due to their size, the figures (figures 11, 12, 13) are placed after the conclusion of this report. The summary of the results are that all three models' predictions worsen for larger values. This tells me that I should investigate using only the logarithmically transformed new cases per million, which would rescale the data in a way to equalize the distribution of values. At the very least, all three models performed better than the surprisingly accurate naive baseline, which indicates to me that I did not make critical blunders.

For the amount of effort, the best option appears to be the ridge regression model; it not only performs the best in terms of the loss function, it also took the smallest amount of time to implement and train. The one caveat, however, is that I would want to see whether this difference in performance generalizes to other forecast intervals, i.e. longer than a single day. For future work I also would like to explore other changes such as: changing the subset of the data which is used to train the model. What I mean by this is to use only the frames which have a substantial amount of new cases per million data, or I could truncate the dates such that I only use data after the peak spread of the pandemic. Alternatively I could include only the time frames in which all countries have one confirmed case. The downside to these operations is that they reduce the already small number of samples. Another method of data manipulation is to explore the inclusion of different amounts of features (which I have performed but is hard to quantify succinctly) or even the inclusion of different numbers of countries for more targeted models.

In conclusion, the problem is obviously a very dynamic and difficult one. Due to the scale, capturing all information relevant to the pandemic is impossible, especially when it is a quantity which is hard to quantify such as human behavior. I believe the most important part is to somehow quantify human behavior in a better way than the government response index. The

reason is that not everyone obeys the government and so there is the “official” stringency value and an “effective” stringency which likely is more accurate in its representation of the actions of a country’s citizens. I believe that this project and the many others like it will in fact remain relevant for years to come, because COVID-19 is not going away anytime soon. I suspect that it will be much like the weather and other travel advisories where forecasts and recommendations are provided as guidelines for daily behavior.

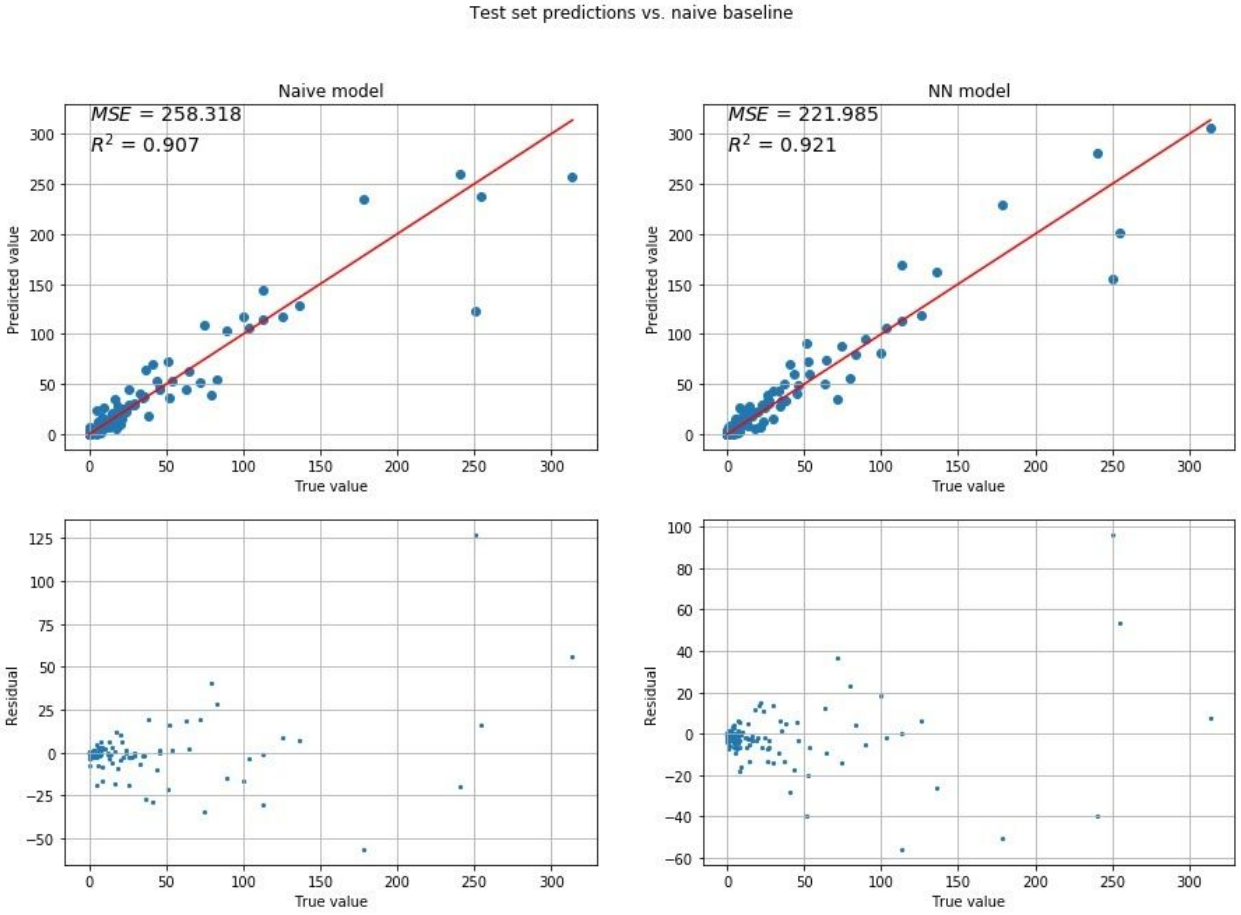

Figure 11: The performance analysis visualized for the fully connected neural network model.

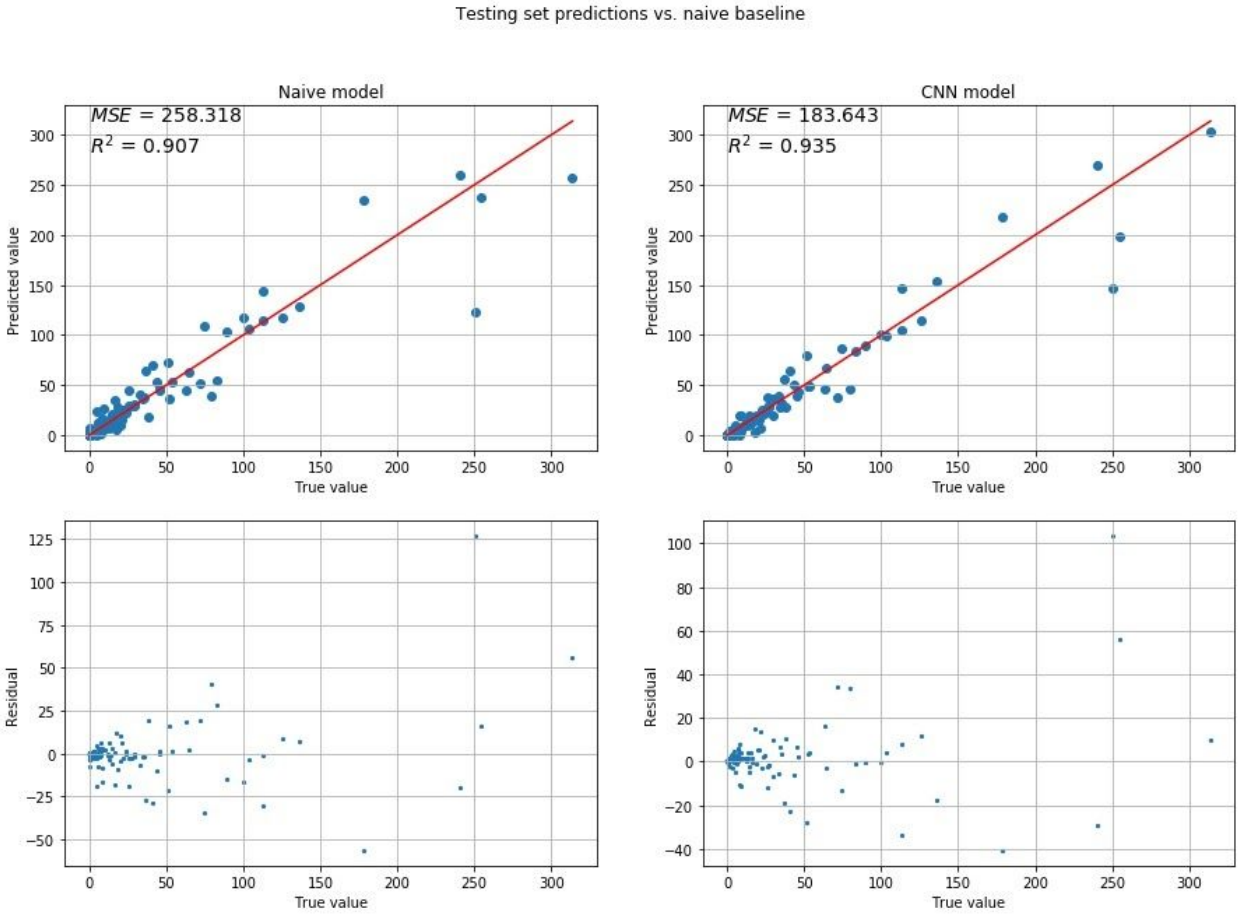

Figure 12: The performance analysis visualized for the convolutional neural network model.

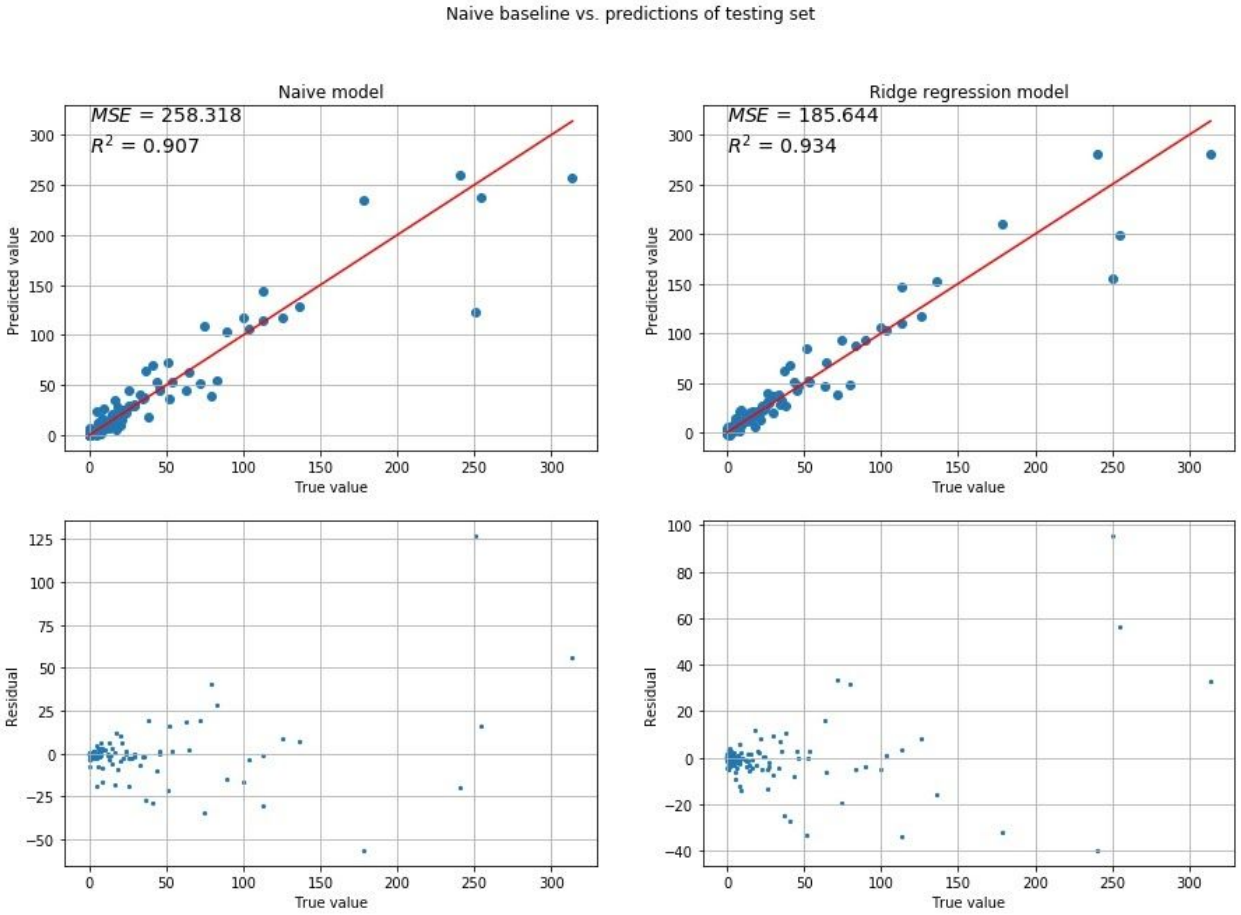

Figure 13: The performance analysis visualized for the ridge regression model.