Using Machine Learning to Predict Covid-19 Cases and Deaths

Team: Diego Diaz: diegodiaz, Raymond Eid: reid7, Piyush Tank: piyushtank

Executive Summary

As the virus SARS-CoV-2, commonly known as Coronavirus, which causes the disease COVID-19, has become a global pandemic; leaders at all levels of government, municipal, regional, and state, have had to make important policy decisions, often with very limited information. While many questions remain about the nature of the virus regarding the infection rate, the mortality rate, and the percentage of the infected individuals who are asymptomatic, a significant amount of research aiming to improve the policy response to the pandemic is being carried out. In the past few months we have observed high variation in the spread of the virus amongst countries of the world. Some countries have been turned upside down by the virus, initially Spain and Italy, now the United States, Russia, and Brazil, while others seem to have contained the virus in a swift, systematic manner, e.g., South Korea, Taiwan, New Zealand.

Governments have considerable discretionary power to implement different degrees of social-distancing measures and strengthen the public healthcare system to help contain the spread of the virus. Given the high mortality rate of the disease, 5.7%[1] worldwide, the spread of SARS-CoV-2 has become the biggest policy issue in the world during the past decades. Causing to this day, June 10, as many as 404 thousand deaths worldwide, and a worldwide economic recession with record unemployment rates.

In this paper we describe our methodology and results in attempting to predict the number of confirmed cases of COVID-19 and the number of deaths caused by the disease based on data of public policies and other country-level economic variables. As we try to fit a continuous variable, we treat our objective as a regression problem. We fit three types of supervised models: Linear Regressions, Decision Tree Regressions, and Random Forests. As a robustness test we fit a Multilayer Perceptron by treating the problem as a classification problem.

We find that the regressions based on trees, single decision tree and random forest, lead to better predictions when trying to predict both cases and deaths, with the single decision tree performing better than all other models by our chosen metric, the root mean squared error (RMSE). We obtain an RMSE of 25,786 when predicting confirmed cases with our decision tree, in comparison to 26,285 with the Random Forest regression. In comparison, our linear regression models range from 95,000 to 200,000 in the same metric.

The policy implications of accurately predicting cases and deaths are considerable since social-distancing measures have a drastic impact on the economy. Because of this, taking policy action from overestimated predictions can be extremely costly for governments, an idea that we explore in this paper in more detail as we

[1] Although the mortality rate is not clear at the moment since many of the infected are asymptomatic, we calculated it based on the total number of deaths divided over the total number of cases. The real rate is likely lower than this and varies between countries with the healthcare system capacity and overall health of the population.

estimate which policies are more effective to decrease the number of cases and deaths.

Background and Overview of Solution

The policies that countries have chosen to implement before and after their initial outbreaks seemingly have direct causal impact in the number of confirmed cases and resulting deaths. As virus containment measures have varied greatly across countries, it has become very hard to discern the optimal set of policy measures for a given country to pursue. Furthermore, a country's preparedness for a pandemic is acknowledged to be a crucial determinant in the importance of policies to contain the spread of the virus. We will define these measures of preparedness in the Data section, as well as government responsiveness in terms of specific policy measures.

It is hard to discern from the data what the "best" course of action is for a country to take, and national priorities must be taken into account - wealthier countries can more easily sacrifice national economic growth and welfare for strict lockdown measures than lower income countries. Writing for Foreign Policy[2], Yale professor and postdoc Mobarak and Barnett-Howell address the reasoning behind the variation in countries' priorities and its policy implications. They state, "social distancing interventions and aggressive suppression, even with their associated economic costs, are overwhelmingly justified in high-income societies," whereas "imposing strict lockdowns in poor countries—where people often depend on daily hands-on labor to earn enough to feed their families—could lead to a comparable number of deaths from deprivation and preventable diseases." In the pair's forthcoming study (2020), they break down the cost-benefit trade off of social distancing measures in dollars, concluding that "the economic value generated by equally effective social distancing policies is estimated to be 240 times larger for the United States, or 70 times larger for Germany, compared to the value created in Pakistan or Nigeria."

We sought out to make a model that would be useful for considering this tradeoff, one that can inform decision makers at a national level about the expected spread of the disease based on a menu of policies. The policy metrics used in our models, discussed in the Data section, include school and workplace closing, stay at home requirements, Covid testing, and contact tracing. An individual using our models can fine tune policy measures to be stricter, for example, and subsequently observe the forecasted number of confirmed cases in the country roughly one week into the future. The models can also be used to forecast future cases without any change in policy.

Our models are intended to provide insight to leaders, but also are useful in informing the public about the importance of lockdown measures. Ultimately, our hope is that policy making in the age of Covid-19 will be based on empirical evidence rather than gut feeling due to the availability of forecasting tools such as our models.

Data

We used three sources of data, two of which were on a country-date level, to train and predict with our model. Before any feature engineering to encode countries into dummy columns, our dataset consisted of 39 features.

For statistics about Covid cases and deaths on a country-day unit, we used data from the Johns Hopkins Coronavirus Resource Center, which has been tracking the spread of the virus for 188 countries with data

[2] https://foreignpolicy.com/2020/04/10/poor-countries-social-distancing-coronavirus/

ranging from January 22, 2020 to the current day. We used these two variables as our outcome label.

Our second source of data, crucial in the policy analysis, is Oxford University's Covid-19 Government Response Tracker, which also contains data on a country-day unit ranging from January 1, 2020up to the current day. The dataset contains 8 containment and closure measures columns, 4 economic measures, and 5 health system measures, which are either categorical or continuous type, fleshed out in Appendix I. The dataset also included indices generalizing each category, encoded as Government Response Index for containment/closure measures, Containment Health Index, and Economic Support Index. Stringency Index aggregates all three indices into one.

Lastly, we use coronavirus-relevant data from the World Bank to control for the preparedness of each country to handle an outbreak. The data are in country-year units and thus do not vary across time in our dataset. In addition, for countries who do not have data for a specific metric in 2019, the most recent year with information is used. Variables and their units are described in the table under Appendix I.

Derived from the Johns Hopkins date, cases, and deaths columns, we engineered the following features: Day Count, Days Elapsed Since First Case, Daily New Cases, and Daily Deaths.

Data cleaning involved standardizing country names, accomplished via a renaming dictionary in Python and compiled via iterative comparisons between country names in the three datasets. In retrospect, the task could have been accomplished more efficiently by merging on country code, however, the Johns Hopkins dataset was missing this column. After cleaning country names, we merged the JHU and Oxford data on Country and Date via an inner join, with special attention given to information loss. We determined that we would have to sacrifice training our model using the following countries due to inconsistencies in the two datasets: Palestine, Kosovo, Dominica, San Marino, Guam, Andorra, Greenland, Gibraltar, Bermuda, South Sudan, and Somalia. We also lost information on various Chinese states, including Taiwan, which is upheld as an exemplary case of proper State management of the coronavirus and would have been a useful case for our model. Finally, the World Bank data was merged on the larger dataset via an inner join on the Country columns.

We encountered missing values from the Oxford Policy dataset, occurring in the most recent week of data, so we decided to drop the last week of rows for each country. This information loss is somewhat limiting but was deemed to be better practice than filling the rows with the most recent observation's values, for we could not simply assume policy measures are not changing.

Visualizations:

Figure 1) South Korea and New Zealand, and to a less extent Lebanon[3], have received media attention for controlling the spread of the virus. Notable features of the graphs is that new cases never exceed 100 for New Zealand and Lebanon (albeit they are relatively less populous countries), and there is great variation in each country's stringency index, suggesting there is no one-size-fits-all prescription of policy measures for a country to adopt.

Figure 2) A variety of countries' logged confirmed cases vs Government Response Index. Generally Gov Response increases with cases, showing that governments are indeed reactive to outbreaks. One interesting trend is New Zealand (the yellow line) which has experienced a significant drop in Government Response as

[3] https://www.washingtonpost.com/world/middle_east/lebanon-is-in-a-big-mess-but-on-coronavirus-its-doing-something-right/2020/04/21/a024496a-83e0-11ea-81a3-9690c9881111_story.html

they no longer have any confirmed cases in the countries and have relaxed almost all containment measures.

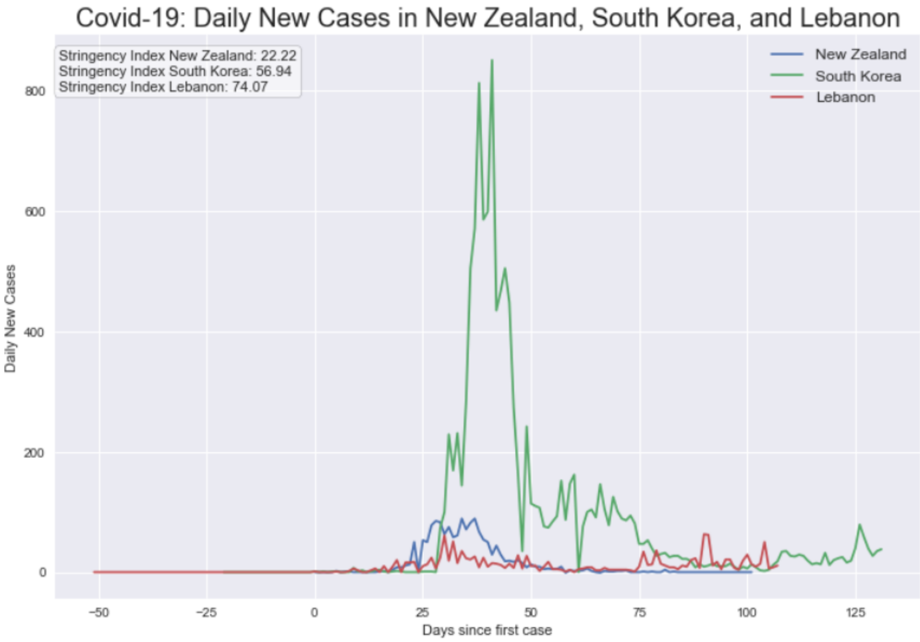


Figure 1

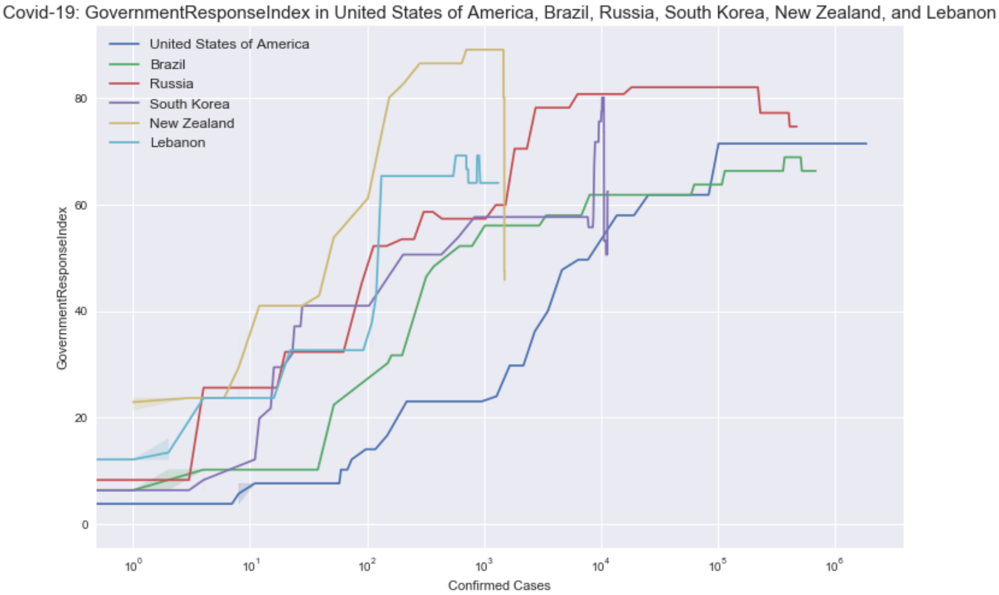


Figure 2

Machine Learning and Details of Solution

Since our target variable (e.g. number of confirmed cases of Covid-19 in each country) is a continuous variable ranging from zero to hundreds of thousands, we are dealing with a supervised machine learning problem in which we require a regression method to make predictions. We use country-day observations from a total of 153 countries from January 1st, 2020 to June 9, 2020, to implement several regression methods, including linear, ridge, lasso, decision tree regression, and random forest regression.

We first approach the machine learning problem by focusing on finding a model with the maximum predicting power when using our features, for which we originally have a total of 34 when combining the Oxford and World Bank datasets. Since 14 of these features are of categorical type (see the list in table A.1 in the appendix), we perform one-hot encoding on them, which increases the number of features to 215.

We follow a standard machine learning pipeline to fit the models. After one-hot encoding the categorical variables we still need to pre-process the rest of the features by the process of scaling, which in our case consists of demeaning and dividing by the standard deviation so that every one of these 20 features has mean 0 and standard deviation of 1. We do this to directly compare feature importance afterward when performing linear regressions. The last step of the pre-processing part of our pipeline involves the target variable, either confirmed cases or deaths. When fitting a linear model, we require the data generating process to be linear on the independent variables, but given the nature of an infectious disease, as it spreads from humans to humans by being in proximity, the number of cases each day is proportional to the number of cases yesterday. In other words. The growth of the target is exponential; therefore, we take its logarithm as our target when performing linear regressions. Since we have many zeros in the data, we add a small number before taking logarithm.

The next step in our pipeline is splitting the data in training and testing sets. Since we are dealing with panel data and our objective is to predict the outcome in the future, we opt to split the data in the same date for every country, leaving every observation to the future of that day in the test set and every observation in the past and until that date, in the training set. We arbitrarily choose this date as two weeks from the day of the last observation available. Therefore, the last two weeks of observations from our data is the test set in which we evaluate our model performance.

For Cross-validation, we use the Time Series Nested Cross Validation approach, where the different folds of data are not determined randomly but with each fold is increasing in time sequentially.

After splitting the data it remains to fit the models to the training set. We measure the performance of each model by calculating the Root Mean Squared Error (RMSE), as it gives us a measure of accuracy which is easily interpretable as it can be measured in the same units as the target variable, which is beneficial when explaining the model to a less technical audience like policy-makers.

Linear Regression model is a first choice for such prediction problems. But a simple linear regression assumes a parametric model of the data generation process and is prone to overfit the training data when there are many features. While Lasso and Ridge are effective regularization techniques to reduce the overfitting in Linear Regression they still impose a parametric form on the data. Decision tree Regression method helps relaxing the parametric assumption and thus increasing the accuracy of the model. And the Random Forest Regressor helps reduce the overfitting of training data. We have trained and tested all these Machine Learning models to choose the best model for prediction.

As a robustness test, we considered a different approach by treating the problem as a classification problem and fitting a neural network to the data, classifying every future country-day observation to a number of cases previously found in the data. We do this to compare performance and measure the accuracy of predictions under a different class of model. We implemented a Multilayer Perceptron (MLP) with one hidden layer with 100 nodes, whose output classifies the outcome into 4627 possible values, which is expected as this is equal to the number of different values our target takes in the training set. By fitting the activation function of each node to a rectified linear unit function (RELU), we obtained the metrics to compare performance with our regression methods. Although an MLP classifier has not been applied in the literature for the problem at hand that we know of, it has been used to the determine the status of COVID-19 patients (Al-Najjar & Al-Rousan, 2020) and to fit the parameters of the differential equations that determine the dynamics of a pandemic (Dandekar & Barbastathis, 2020), known as the SIR (Kermack & McKendrick, 1991) and SEIR model (Fang et al. 2006).

Evaluation and Results

While evaluating models, we use the Root Mean Squared Error (RMSE) metric and choose the model with the lowest RMSE. For Confirmed Cases and Deaths, RMSE for different models is as below[4]:

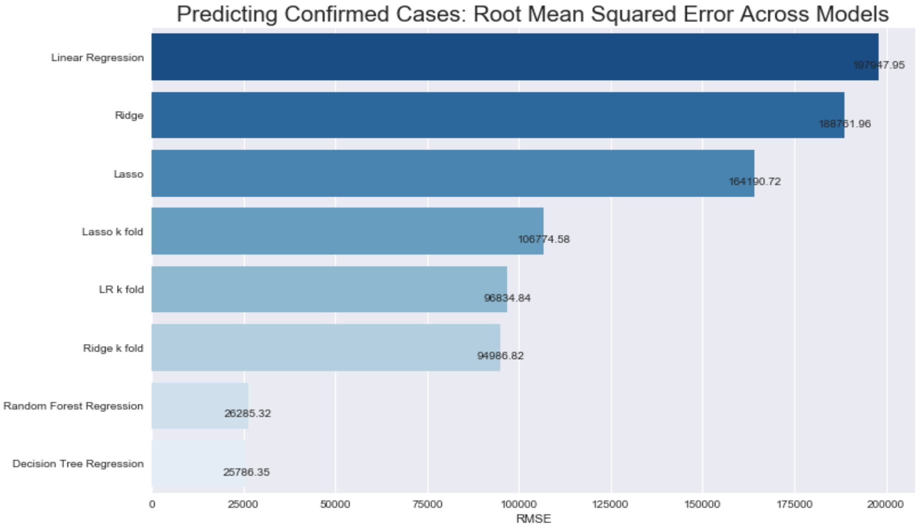


Figure 3

[4] For robustness we also fitted a Multilayer Perceptron neural network when treating the problem as one of classification, obtaining a RMSE of 28834 when predicting confirmed cases and 11058 when predicting deaths.

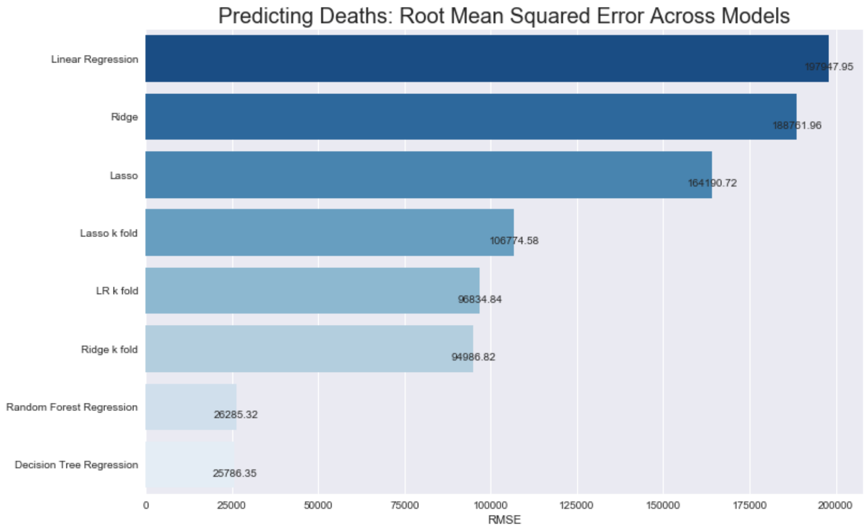


Figure 4

As it shows that for Decision Tree Regressor and Random Forest Regressor the RMSE is significantly lower than other models. In Linear Models, for Confirmed Cases, Lasso Regression produces lower RMSE than Ridge and Linear Regression. Thus, indicating a little bit of overfitting with the Linear Regression model. For Deaths, Linear Regression produces lower RMSE than Lasso and Ridge, thus indicating that for Deaths as an outcome, we are still underfitting the data and would require to get more features to reduce RMSE. Although the Linear Models are not producing the lowest RMSE, the comparison among them provides information on Bias and Variance tradeoff. Decision Tree Regressor and Random Forest Regressor, when allowed to use all features as layers and produces the most accurate Confirmed Cases and Deaths. We have also compared the prediction with the Neural network with Linear Regression.

Below are visualizations of model predictions vs real outcomes, using both a Neural Network and a linear regression in Kenya and Austria. The two countries were chosen as they are illustrative of a general trend we observed - at times the regression outperformed the neural network and vice versa. Generally speaking, upon examination, the visualizations showed the Neural Network more accurately reflecting real outcomes, i.e., lower error.

Figure 5: Comparing prediction of Linear Regression with Neural Network model for Kenya

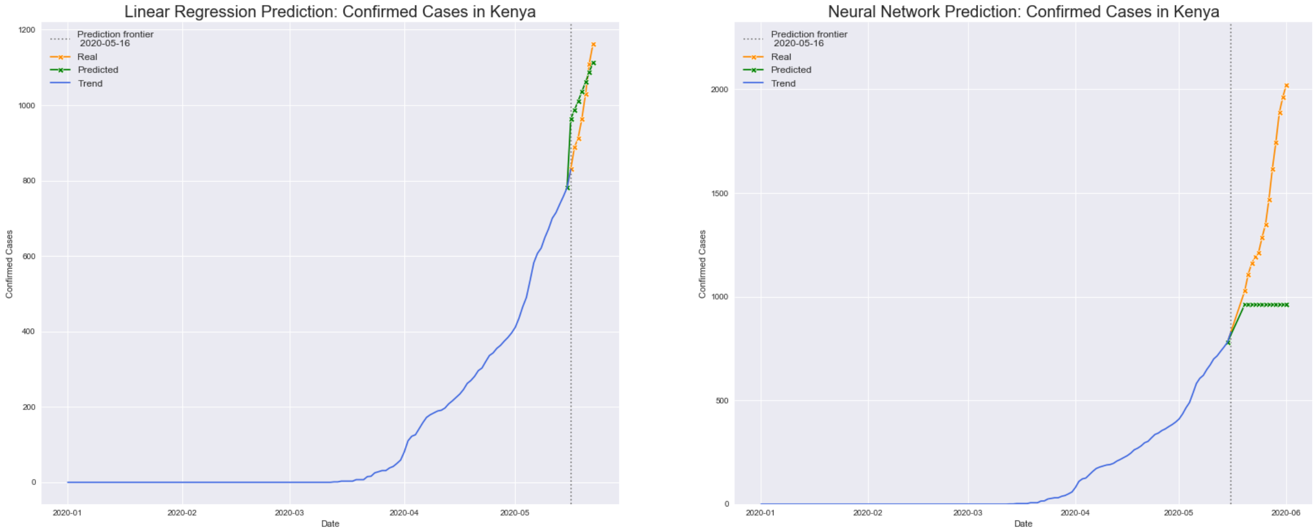
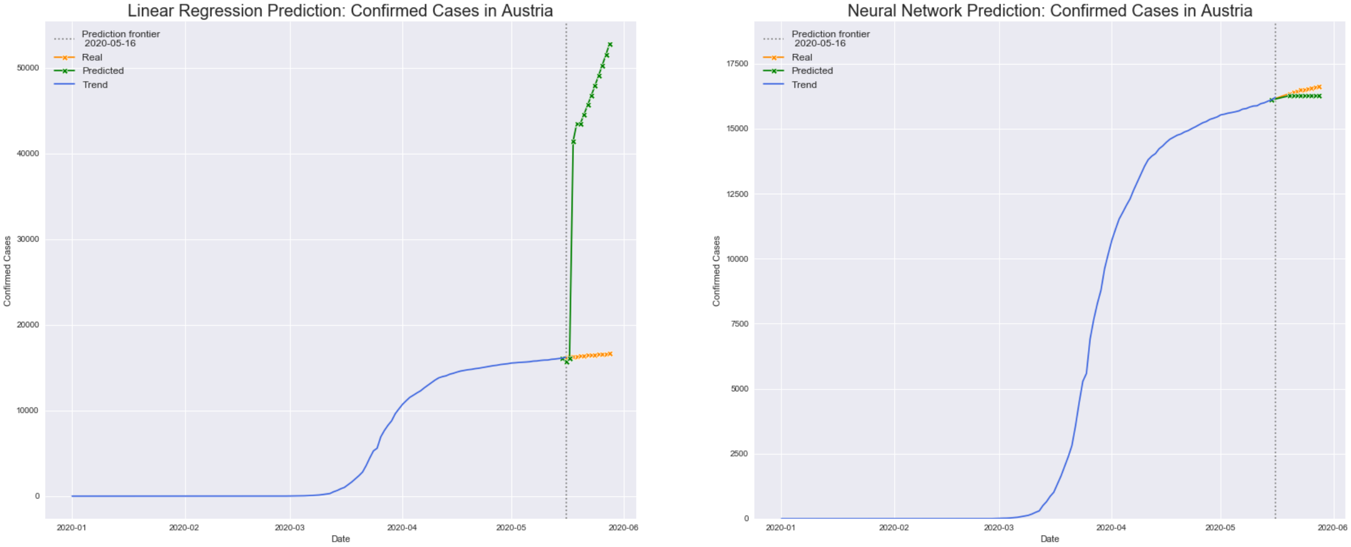
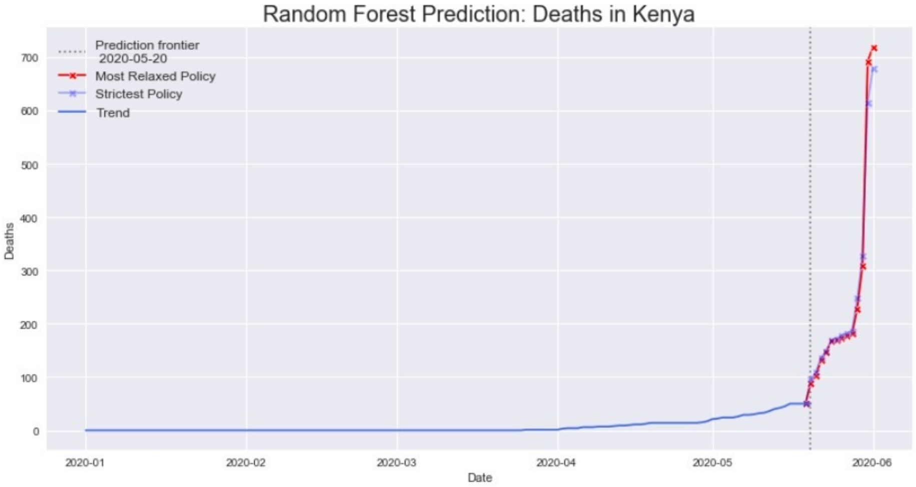


Figure 6: Comparing prediction of Linear Regression with Neural Network model for Austria



Policy Recommendations

National governments are facing a difficult policy situation of reducing the impact of Covid-19 with minimum economic impacts due to lockdowns, they need to figure out the optimal lockdown policy given the country's situation since the first Covid-19 case has been registered. Our model can help make this decision, by comparing the potential outcomes in Confirmed Cases and in Deaths by various lockdown strategies, and although it would benefit from refinement, the random forest classifier below shows that stricter policy measures will lead to fewer deaths.



Ethics

There have been reports from major newspapers about some countries underreporting their cases and deaths of Covid-19. Having under-reported values of our target variable can lead to downward bias in regression estimates, which is an important concern when predictive models are used in policy making. Expecting lower cases and deaths may lead a policy maker to not anticipate the needs of the public healthcare system and to void or decrease the intensity of preventive measures as they come at an economic cost. At the same time, decreasing the intensity of these policies will cause cases to increase. If the expected number of cases is below the real value, cases and deaths will rise much higher considering the exponential nature of an infectious virus like SARS-CoV-2. For this reason, and the high value of human life, having downward bias in estimates is likely more costly than having upward bias. To account for this issue, as a robustness test we opt to train a linear regression model without three countries that have been considered to underreport COVID-19 cases by the New York Times[5] [6] [7]

We find no differences in the average RMSE of cross validation obtained by the model when leaving out China, Russia and Brazil. However, underreporting can be a more systematic issue if it is happening already for these countries, so it is an ethical concern and we recommend a thorough look into data gathering and

[5] A Coronavirus Mystery Explained: Moscow Has 1,700 Extra Deaths. NYT, May 11, 2020. https://www.nytimes.com/2020/05/11/world/europe/coronavirus-deaths-moscow.html

[6] China Raises Coronavirus Death Toll by 50% in Wuhan. NYT, April 17, 2020. https://www.nytimes.com/2020/04/17/world/asia/china-wuhan-coronavirus-death-toll.html

[7] Furious Backlash in Brazil After Ministry Withholds Coronavirus Data. NYT, June 8, 2020. https://www.nytimes.com/2020/06/08/world/americas/brazil-coronavirus-statistics.html

reporting mechanisms of countries before making policy decisions based on predictive models.

Limitations, Caveats, Suggestions for Future Work

Given our model estimates targets such as the number of infected people and number of deaths, a suggestion for future work is to compare the model accuracy with a theoretical SEIR model (suceptible-exposed-infected-recovered) as in Dandekar & Barbastathis (2020). The original SEIR model was developed by Fang et al. 2006, and consists in separating the population in 4 groups and modeling their behavior with a set of differential equations that determine the dynamics of the pandemic. Such parameters can be fitted with the data for a given geographic area.

Given the biases of the confirmed cases discussed in Section 8 on Ethics, in future work we could create an instrumental variable for confirmed cases, such as rate of covid-19 cases in pregnant women as some researchers have already done. Pregnant women are the most likely to be tested and the accuracy of their tests is of special importance.

References

Al-Najjar, H., & Al-Rousan, N. (2020). A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea. Eur Rev Med Pharmacol Sci 2020; 24 (6): 3400-3403

Barnett-Howell, Z., & Mobarak, A. M. (2020). Should Low-Income Countries Impose the Same Social Distancing Guidelines as Europe and North America to Halt the Spread of COVID-19? *Yale School of Management*, 1–7. Retrieved from https://som.yale.edu/sites/default/files/mushifiq-howell-v2.pdf

Dandekar, R., & Barbastathis, G. (2020). Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. *medRxiv*.

Fang, H., Chen, J. & Hu, J. 2006 Modelling the sars epidemic by a lattice-based monte-carlo simulation. In 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 7470–7473. IEEE.

Kermack, W; McKendrick, A (1991). "Contributions to the mathematical theory of epidemics – I". Bulletin of Mathematical Biology. 53 (1–2): 33–55. doi:10.1007/BF02464423. PMID 2059741.

Kniesner, T. J., & Viscusi, W. K. (2019). The Value of a Statistical Life. *Forthcoming, Oxford Research Encyclopedia of Economics and Finance*, 19-15. DOI: 10.26355/eurrev_202003_20709

Appendix I.

Table 1: Oxford University Covid-19 Government Response Tracker

Name	Description	Coding
Containment and closure measures		
C1_School closing	Record closings of schools and universities	0-3
C2_Workplace closing	Record closings of workplaces	0-3
C3_Cancel public events	Record cancelling public events	0-2
C4_Restrictions on gatherings	Record limits on private gatherings	0-4
C5_Close public transport	Record closing of public transport	0-2
C6_Stay at home requirements	Record orders to "shelter-in-place" and otherwise confine to the home	0-3
C7_Restrictions on internal movement	Record restrictions on internal movement between cities/regions	0-2
C8_International travel controls	Record restrictions on international travel	0-4
Economic measures		
E1_Income support (for households)	Record if the government is providing direct cash payments to people who lose their jobs or cannot work.	0-2
E2_Debt/contract relief (for households)	Record if the government is freezing financial obligations for households (e.g., stopping loan repayments, preventing services like water from stopping, or banning evictions)	0-2
E3_Fiscal measures	Announced economic stimulus spending	Continuous (USD)
E4_International support	Announced offers of Covid-19 related aid spending to other countries	Continuous (USD)
Health System Measures		
H1_Public information campaigns	Record presence of public info campaigns	0-2
H2_Testing policy	Record government policy on who has access to testing	0-3
H3_Contact tracing	Record government policy on contact tracing after a positive diagnosis	0-2
H4_Emergency investment in healthcare	Announced short term spending on healthcare system, e.g., hospitals, masks, etc.	Continuous (USD)
H5_Investment in vaccines	Announced public spending on Covid-19 vaccine development	Continuous (USD)

Table 2: World Bank Variables

Variable	**Units**
General features	
Gross Domestic Product	USD
Population	Individuals
Life Expectancy at birth	Years
Relevant Covid-19 features	
Physicians	per 1,000 individuals
Universal Health Care Service Coverage Index	Continuous 0-100
Diabetes prevalence Ages 20 to 79	Percent of Population
Current Health Expenditure Per Capita.	USD, PPP