Ludwig-Maximilians-Universität München
Department of Economics

Machine Learning: Applications in Economic Research
Project of the team Robocode

"Which countries are more affected by COVID19? Application of machine learning methods to find out the behind story"

Supervisor: *Dr. Nadzeya Laurentsyeva*

Authors:
Anastasia Matveeva
Berkay Ölken
Florian Schuchardt
Hajar Mohd Razali

Winter Term 2020/21

March 15, 2021

CONTENTS

INTRODUCTION 3

DATA 4

METHODOLOGY 5

RESULTS 7

CONCLUSION 8

REFERENCES 9

APPENDIX 10

INTRODUCTION

The world has been struggling with many problems such as earthquakes, tsunamis, wars, scarcity of resources, other natural events and health diseases. Humanity took its lesson from each of these events; they built stronger constructions against earthquakes, they found better ways to minimize the impact of tsunamis, optimal usage of resources is still increasing its popularity to deal with scarcity. Science and other scientific disciplines have been developed immensely to overcome the problems. The vast and fast reaction of scientific solutions were successful against most of the above troubles because they were evident, stationary and relatively predictable to be taken proper cautions. Yet, what about health diseases that affect mass people? The three biggest pandemics throughout history caused the loss of 300 million peoples' lives[1]. The main reason is because these diseases are non-stationary since there are an infinite number of possibilities that causing-microorganism can mutate, which basically prevents scientific solutions to respond quickly. Today, humanity has again found itself in the hearth of a pandemic crisis, which is called COVID-19. The name originated from CoronaVirus Disease in the year of 2019. As Plague, Swine Flu and Ebola, this virus also mutated and spread to humans from animals. So far until its first year, it took around 2.5 million people's lives and created a deep shock for others. Even if there was no quick scientific solution in the beginning, the vaccines are now able to be developed after around a year. Yet, this will not reverse the damage that has been done.

Even if the deaths are the most tragic outcome of pandemics, there are other consequences which have strength to distort the fundamentals of human lives. As a domino effect, pandemics are followed by the economic consequences such as increasing unemployment, inflation, debt crisis, decreasing reserves and investments. Periods such as COVID19 leads us to observe in what sense and how dependent the countries are as well as their economic power to compensate for the negative effects of the problem. Developed countries introduced their economic stimulus packages one-by-one against the COVID19 and today, in total $10 trillion worth of stimulus, which is around 15% of world GDP, is pumped through the economic system by all countries.[2] This is mostly led by super-powers such as the United States and Germany, yet it signals the severity of the problem and its possible upcoming results. Furthermore, it is quite challenging to understand what the economic responses of different countries would be, considering that most of them have different dependencies, economic

[1] https://www.visualcapitalist.com/history-of-pandemics-deadliest/

[2] https://www.mckinsey.com/industries/public-and-social-sector/our-insights/the-10-trillion-dollar-rescue-how-governments-can-deliver-impact

structures and abilities. In the light of this, it is highly important to understand and estimate the economic exposure of the COVID19 to the countries with divergent economic structures which would serve as an early warning for the future periods.

This paper is aimed to introduce an efficient strategy of using Machine Learning Methods (MLM) to understand the vulnerability of countries against the COVID19 shock. The usage of ML methods stands substantial in two ways: First, considering the mentioned economic consequences (and more) as well as the variability of economic composition of countries, it provides efficient solutions to deal with such high-dimensional problems. Second, given the abundance of nonlinear variables and highly volatile environment, they provide optimal ways to capture the sweet-point for the possible bias-variance relation. Among the ML methods, Random Forest (RF) and LASSO (Least absolute shrinkage and selection operator) were tested against an alternative OLS method. Besides the vulnerability analysis mentioned above, this study is also reporting the performance differences of these three methods.

DATA

The data of this study is mostly based on World Bank Databases[3], there are 11 macroeconomic variables used in the analysis and additional counts such as number of deaths and number of cases are taken into account for the COVID19 assessment which is downloaded from "Our World in Data" source[4]. The data for tourism and travel activities of the countries are taken from WTTC[5]. There are in total 209 countries and the quotation of all the variables as defined below based on their values for the year of 2019. The main aim here is to be comprehensive and accurate by using leading macroeconomic variables through the analysis. In that sense, the variables can be classified into two groups based on the information they carry: 1- variables representing the dependency of countries and 2- variables related with the economic power of countries. The variables in the former are defined as: foreign direct investments inflow (%GDP), received remittances (%GDP), food imports (% of merchandise exports), fuel exports (% of merchandise exports), net ODA (% of GNI, official development assistance), ores and metals exports (% of merchandise exports), gross debt (% GDP), travel and tourism inflow (% GDP). The variables in the latter are: unemployment (% of labor force), total reserves in month of imports (number of months can be compensated by only using reserves) and GDP growth rate (annual % change). An additional group can be

[3] https://databank.worldbank.org/home.aspx
[4] https://ourworldindata.org/covid-cases?country=IND~USA~GBR~CAN~DEU~FRA
[5] https://wttc.org/Research/Economic-Impact

defined for the COVID19 variables which directly reports the number of deaths and number of cases per million as of 27 February 2021.

In the original sample, there were 264 countries but given that the setback of a considerable number of variables is missing in the data for some specific countries, they had to be dropped from the sample. Yet, not all the countries suffered from the high number of missing data problems, some of them had few missing variables which could still play a role in the analysis. In that sense, a prediction procedure is applied for those countries to predict their missing variable data by using their complete set of variables. This prediction procedure is basically utilizing the covariance in between missing variables and other variables which are complete. One big advantage[6] of this method is that it provides better and more accurate values considering the other techniques dealing with the missing data issue (e.g. Imputation method, mean/median replacement method).

METHODOLOGY

In our research project we would like to compare the results of three methods that we used: OLS regression, Random forest and LASSO, we have chosen 4 dependent variables and 8 independent variables, all of them are listed in the Data section.

In order to choose the right methods, we start by categorizing the problem. Categorizing it by the input, since we have labeled data and predictor variables, we know that it is a supervised learning problem. On the other hand, categorizing by our outputs, since they are numbers, we know that one of the methods to solve this is by seeing it as a regression problem.

As part of data preparation, we merge datasets from different sources into one dataframe. As for the availability of the data, as stated above, part of the data is missing, so we pre-process, profile and clean it with the method of prediction. In order to find out the missing values, we use the features that do not contain null values. One of the main positive aspects of using this method, the covariance between missing value columns and other columns are taken into account.

Further, we proceed with transformation of dependent variables into log, because their distributions are not normal. Using the logarithm of one or more variables improves the fit of

[6] https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e
Section: Prediction of missing values

the model by transforming the distribution of the features to a more normally shaped bell curve.

After data preparation, we use OLS regression as the baseline model for our prediction. We chose OLS to serve as a performance benchmark for the ML models to predict our outcome variables. It is essential to highlight that because of the simplicity of the OLS algorithm, it works well when we have many features. More complex algorithms may suffer from overfitting from the combination of a huge number of features and small amount of data sets where a simple linear regression may thrive. However, OLS could be unstable in the case of redundant features.

Next, we build a Random Forest model on a first randomly-selected sample from our training data and use five-fold cross validation to identify the optimal number of trees. Generally, RF produces better results, works well on large datasets, and is able to work with missing data by creating estimates for them. However, they pose a major challenge that is, they cannot extrapolate outside unseen data. Next, by using cross-validation we determine the optimal number of trees. (For example, for «GDP growth» n=112) Assuming that the higher the performance, the better dependent variable is explained by independent variable, we come up with that by estimating «log unemployment», «log GDP growth», «log Total deaths» the model performs better with randomization, however «log Total cases» performs worse with randomization.

After Random Forest, we use LASSO regression. In order to find the optimal lambda , we fit LASSO using a five-fold cross validation method. Lasso regression is a type of linear regression that uses shrinkage. This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, as variable selection/parameter elimination (Glen 2015). The motivation of choosing LASSO is that we have to figure out whether there is a relationship between our independent variables and dependent variable. Comparing two methods LASSO and RIDGE, we agreed on LASSO, as we have less significant parameters than RIDGE. We would expect RIDGE to perform better if we would have more equal and significant parameters. Yet, since this is not the case for our structure, LASSO is expected to be a more accurate candidate.

Finally, we evaluate out-of-sample performance of the three models by using five-fold cross validation to compare the predictive powers of the models. The K-fold cross-validation method evaluates the model performance on different subsets of the training data and then

calculates the average prediction error rate. LASSO model turned out to be the most inappropriate for our research among the three.

RESULTS

Inspired by the kaggle dataset on Economic Impact of Covid19 on individuals we agreed that our independent variables are the factors that can explain COVID19 economic exposure. So based on that, we try to find which dependent variable explains our independent variables the best. As the results suggest, unemployment seems to give us the best predictor variable that can explain our inputs. Based on Table 1, 50% of the R-squared scores lie between 0.196293 and 0.344192. The model has a mean score of 0.27326 with standard deviation of (+/- 0.20625). The other variables have less convincing results. For GDP growth, random forest is the best performing model. Based on Table 2, 50% of the R-squared scores lie between 0.051399 and 0.153471. The model has a mean score of 0.12333 with standard deviation of (+/- 0.24804). For Total Cases from Figure 3, the result is inconclusive regarding which model performs the best, as even though for random forest, the median of the scores is the highest in comparison of the other 2 models, the variability of the scores is very high as from figure , we can see that it has a large interquartile range. This could potentially indicate that the scores have a high variance and might be unreliable. On top of that, the median may not gauge its central tendency well. For Total Deaths in Figure 4, the result is inconclusive regarding which model is the best as even though for OLS, the median of the scores is the highest in comparison of the other 3 models but the scores are either negative or hover around 0 which indicates that the features that we have cannot explain our predictor variable well or only very weakly.

From these analysis, we concluded that unemployment is the only variable that can best be used to predict our inputs. Thus, we created the rank of the countries using this variable in Table 5. The most vulnerable countries from our sample according to Unemployment rate in 2019 are African countries, number 1 is South Africa with unemployment rate of 28.47%, 2nd place is West Bank and Gaza with 25.34% , 3rd is Lesotho with 23.86%, 4th is Eswatini with 22.24% and fifth place is Namibia with 19.75%. We are also aware of the probability of underreporting and statistical issues as measurement errors that may occur in different countries, especially in developing countries. We also take the fact into account that the score for our best predictor variable is also weak as it is below 0.5. Perhaps there are better models

suited best for our data or there are other more suitable independent/dependant variables that can be used to address the research question..

As for the comparison of three methods used, we assume that for the dependent variable "Unemployment" and "Total cases" the best fit model is Random Forest. The main reasons for these results could be because of RF's ability to deal with non-linear variables under a possible volatility environment. Given that we have explanatory variables from different characteristics to explain the COVID19 shock, RF might be performing well in some cases. However we cannot use RF for "Total Deaths". The LASSO model performs worse among all these 3 models. As for OLS, it captures higher performance than other models for variable "Total Deaths".

CONCLUSION

In this project, we wanted to get our hands on the real-life data to analyse an important and recent global event: COVID19. We believe it is substantially important to understand the dynamics and impacts of this event, given that it deeply affects the economic structure of countries, independent from their development level. The analysis held considering the fundamental variables which could give clues about both the level of vulnerability of countries and their economic powers. Given the sensitivity of the topic, using Machine Learning methods would be crucial to handle data-rich and complicated analysis processes. Thanks to its efficient variable selection techniques, ability to deal with possible nonlinearities and finding the sweet spot for bias-variance relationship; Machine Learning Methods played a crucial role in this analysis. As above mentioned, Random Forest and LASSO models are tested in this analysis as well as OLS. We believe even if RF does not perform well for all the dependent variables, it was the best performing model among all in overall. In that sense, for further research based on this topic, it is important to highlight the power of RF in this field. Based on our analysis, this project also reports a ranking for the countries with respect to their vulnerability against shocks such as COVID19. This ranking could be utilized in the further policy making processes with a detailed analysis of the economic structure of least and most vulnerable countries.

REFERENCES

1. Almaliki, Z., (2019). Do you know how to choose the right machine learning algorithm among 7 different types? Retrieved from https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60
2. Cassim, Z., Handjiski, B., Schubert, J.,& Zouaoui, Y. (2020) The $10 trillion rescue: How governments can deliver impact. Retrieved from https://www.mckinsey.com/industries/public-and-social-sector/our-insights/the-10-trillion-dollar-rescue-how-governments-can-deliver-impact#
3. Galarnyk, M., (2018) Understanding Boxplots. Retrieved from https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51
4. Gaurav, Y.(2020). Economic impact of Covid19 on individuals. How Covid 19 affected our daily lives. Retrieved from https://www.kaggle.com/gauravmehta13/economic-impact-of-covid19-on-individuals?select=_Survey+on+Economic+Impact+of+Covid19+on++++Individuals++%28Responses%29+-+Form+Responses+1.csv
5. Gavrilova, J., (2020). How to choose a machine learning technique. Retrieved from https://serokell.io/blog/how-to-choose-ml-technique
6. Glen, S. (2015) Lasso Regression: Simple Definition" From statistics HowTo.com: Elementary Statistics for the rest of us! Retrieved from https://www.statisticshowto.com/lasso-regression/
7. Kinha, Y.,(2020). An easy guide to choose the right Machine Learning algorithm. Retrieved from https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html
8. Kumar, S.(2020). Ways to Handle Missing Values in Machine Learning. Retrieved from: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e
9. LePan, N. (2020). Visualizing the history of pandemics. Retrieved from https://www.visualcapitalist.com/history-of-pandemics-deadliest/
10. Roser,M., Ritchie,H.,-Ospina E. & Hasell, J (2020). Coronavirus Pandemic (COVID-19). Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/covid-cases?country=IND~USA~GBR~CAN~DEU~FRA
11. World Bank Data Bank. Retrieved from https://databank.worldbank.org/home.aspx
12. World travel and tourism council Data Bank. Retrieved from: https://wttc.org/Research/Economic-Impact

APPENDIX

Figure 1.

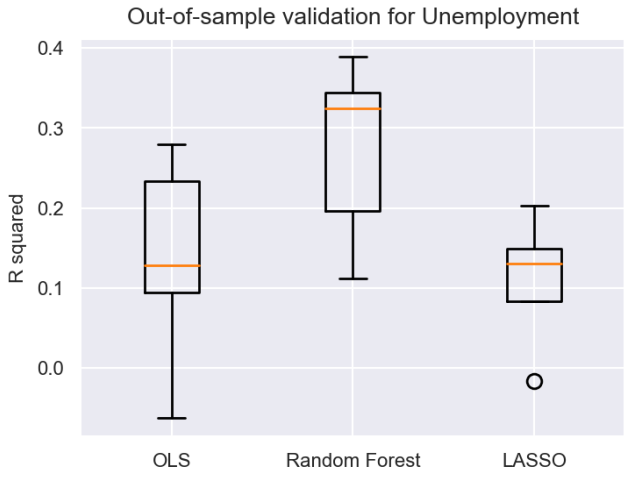


Table 1.

	label	lower_whisker	lower_quartile	median	upper_quartile	upper_whisker
0	OLS	-0.062599	0.094224	0.128297	0.233370	0.279994
1	Random Forest	0.111699	0.196293	0.324867	0.344192	0.389245
2	LASSO	0.083570	0.083570	0.130525	0.149386	0.202668

Figure 2.

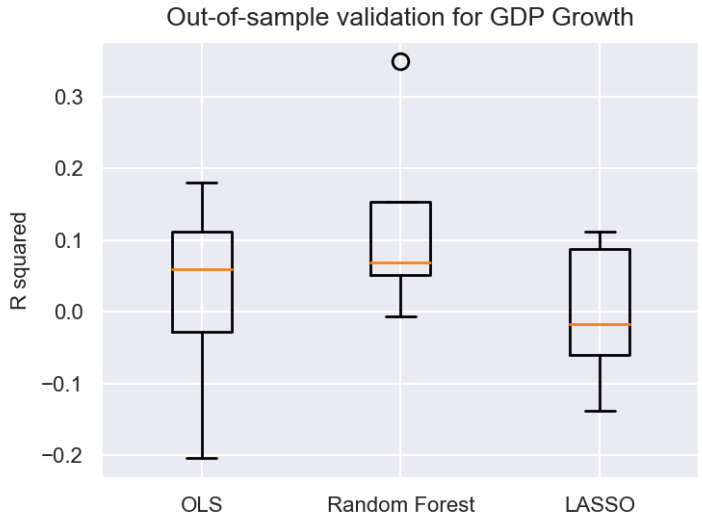
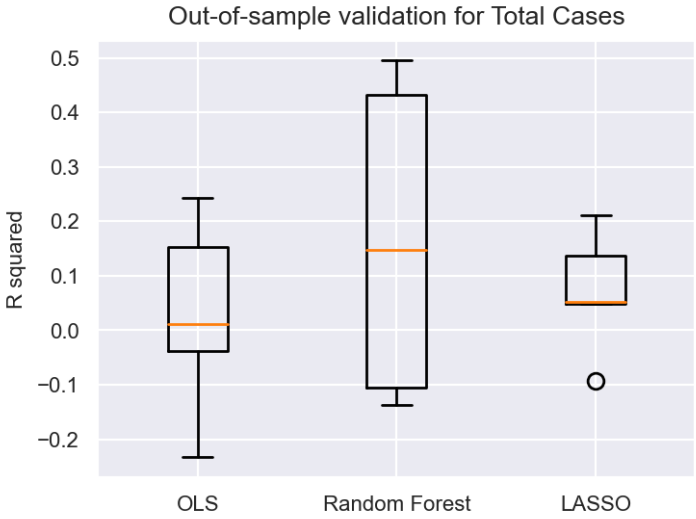


Table 2.

	label	lower_whisker	lower_quartile	median	upper_quartile	upper_whisker
0	OLS	-0.203669	-0.028152	0.060023	0.112149	0.179830
1	Random Forest	-0.006536	0.051399	0.069138	0.153471	0.153471
2	LASSO	-0.137532	-0.059552	-0.017010	0.087671	0.111614

Figure 3.



.

Table 3.

	label	lower_whisker	lower_quartile	median	upper_quartile	upper_whisker
0	OLS	-0.232580	-0.037656	0.010573	0.153287	0.242856
1	Random Forest	-0.137073	-0.105846	0.147326	0.432075	0.495567
2	LASSO	0.048741	0.048741	0.051820	0.137669	0.211381

Figure 4.

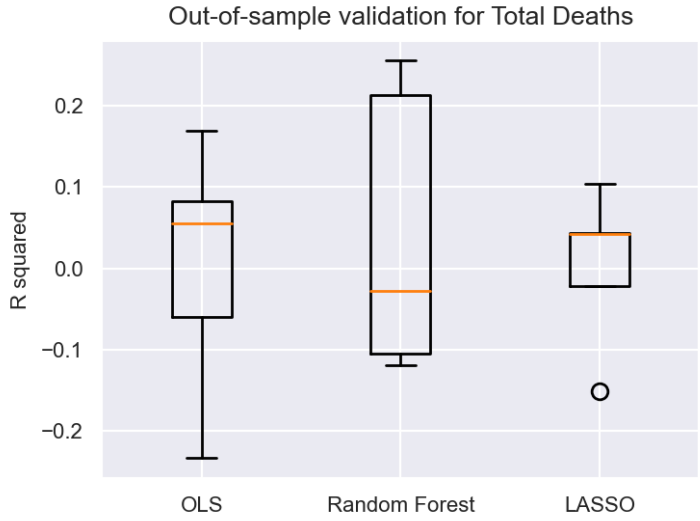


Table 4.

	label	lower_whisker	lower_quartile	median	upper_quartile	upper_whisker
0	OLS	-0.232826	-0.059346	0.055011	0.082559	0.169667
1	Random Forest	-0.118910	-0.104696	-0.027782	0.213611	0.255568
2	LASSO	-0.022352	-0.022352	0.042814	0.042950	0.103580

Table 5.

The 15 most vulnerable countries according to Unemployment rate in 2019

Country	unemployment
South Africa	28.469999
West Bank and Gaza	25.340000
Lesotho	23.860001
Eswatini	22.240000
Namibia	19.750000
Gabon	19.639999
Armenia	18.809999
St. Vincent and the Grenadines	18.620001
Libya	18.340000
Greece	17.309999
North Macedonia	17.260000
Botswana	17.209999
Jordan	16.850000
Sudan	16.760000
Bosnia and Herzegovina	15.690000