Predicting Covid-19 Confirmed Cases and Fatalities for April 2020
Team Name: Ria Chakrabarti

Introduction

This project, in place of a final exam, involved forecasting cases and fatalities from the Covid-19 pandemic for different provinces and countries for the month of April. As part of a Kaggle competition, one of the requirements was to submit a late submission to the competition. The submission was in the form of a csv file labeled submission.csv which includes an id for each prediction as "ForecastId", number of confirmed cases for each prediction as "ConfirmedCases", and fatalities for each prediction as "Fatalities".

Background

Everyday, the current Covid-19 (also referred to as Coronavirus) pandemic is proving to be a puzzling and contentious threat to global health. Politicized agendas combined with a lack of information about how this virus can affect people on a global scale has made it even more difficult for regular members of the public to inform themselves on how to navigate the pandemic.

In order to make informed decisions about which features to include in the data I conducted a small amount of background research based on speculations about the effect age and population density had on the virus.

Age

Early reports of Covid-19 were very clear in stating that elderly people about 80 years of age or older were the demographic with the highest severity of symptoms to the virus (Mahase 2020). As months went by, forecasts made by universities and government agencies started to take age more into account. Demographic information is proving to be critical to governments making policy decisions. A study from the University of Oxford by Dowd et al (2020) asserts that local and national statistics on age demographics seem to affect the projections of Coronavirus fatalities. This study even argues that the structure of population age reflects mortality rates of the virus.

As seen in *Figure 1*, differences in the structure of population age seemingly reflect the number of deaths relative to the population of Brazil and Nigeria. The climate and average temperatures between the nations of Brazil and Nigeria are also relatively similar, indicating further controls in this comparison.

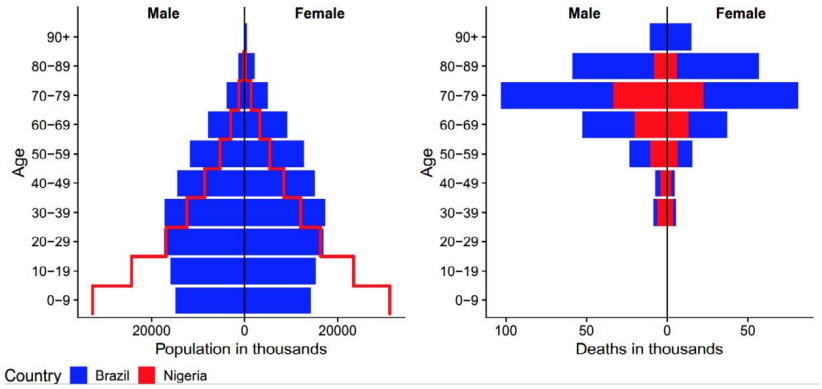

Figure 1: Population age distribution and comparison of deaths due to Covid-19 (Dowd et al., 2020)

Therefore, I chose to include median age by country in the data for the project.

Population Density

There is more contradictory evidence on population density affecting mortality from Covid-19. Population density can obviously affect transmissibility i.e. more people in a smaller space are more likely to come into contact in the first place. However some studies are finding that population density may also have an effect on the severity with which people contract symptoms (Rocklöv and Sjödin 2020). This chart from Our World in Data sourced from the World Bank shows Covid-19 deaths by country by population density of each of those countries and does not show any obvious correlation between population density and deaths.

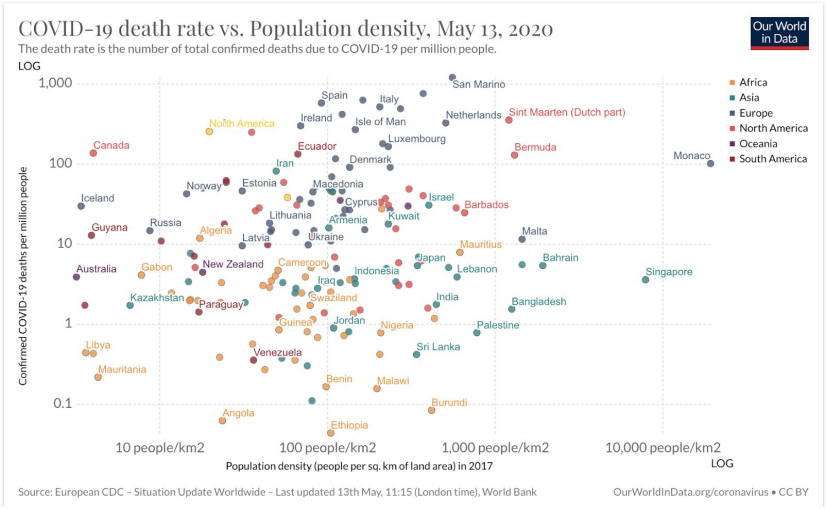

Figure 2: Covid-19 deaths by population density by countries (Global Change Data Lab 2020)

Given the inconclusive data on fatalities by population density and the correlation between the number of cases and population density, I decided to also include it in my project.

Data Preparation

Firstly, based on my background research I chose to include country level data on age and population density. I attained data by country from a source I found on Kaggle in a file called "covid19countryinfo.csv". From this source I merged information into my training and testing data on population density, median age, and urban population on the country level. Then I had to encode the object labels of the given data into integers that could be fed into a regression model with a `LabelEncoder`. This proved to be difficult later on when I tried to use a final model to predict an encoded version of the test set. This is because the given training data is missing over 60 countries that are present in the data set used for submission. Kaggle also only accepts submission files for this competition with exactly 12642 rows, meaning I also could not omit the 60 countries unseen by the training data. I got around this eventually but it provided a lot of difficulty at first.

I used data from the given "train.csv" file for training and validation data. Data from this file before March 19th was used as training data and data between March 19th and March 30th was used as validation data. The given test data which was to be used for submission contained data from March 19th to April 30th.

Model Selection And Evaluation

In order to get a baseline of performance I trained a MultiOutputRegressor with a Support Vector Machine Regressor on just the encoded training data. I chose a Support Vector Machine Regressor mostly because it is based on Support Vector Machines known as a powerful model which incorporates the use of hyperplanes to determine powerful classifications for data while also being simple to implement with a MultiOutputRegressor. I then used Root Mean-Squared Error (RMSE) and Explained Variance Score (EVS) to evaluate the model for a baseline of performance. Root mean square error describes the simple difference between predicted values and actual values. Explained Variance Score can be described in Figure 3 where Var is variance which is the square of the standard deviation, ŷ is the prediction, and y is the correct value. It was found that his method had a RMSE of 6616.44 and an EVS of 2.9959e-09. Explained variance scores can be 1.0 in the best case and lower values tend to indicate worse performance. Clearly, this is a very low score.

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

Figure 3: Explained Variance Score explained by SK-Learn (2020)

I then attempted to train and make predictions only by the Country/Province level. However because each model was insulated to being trained by a specific Country/Province combination, the additional features of population density and median age would have minimal importance because country features were not being compared against each other for training. I used Machine Learning Mastery's "Guide to Multi-step Forecasting" as inspiration to this method of modelling (Brownlee 2018). I created functions that would take models, training data, validation data, and labels as input and train and predict by Country/Province. The average RMSE from using this method with a Support Vector Machine Regressor was 1412.442 and the average EVS was 0.1297877. This is a great improvement from the original baseline evaluations. I also took note differences in the growth of fatalities and confirmed cases in different countries. For example, Chinese data over this time period showed logistic growth while data from Iran represent exponential growth. This inspired me to also use different types of underlying models using the multi-step forecasting method. A LinearRegression underlying model had a negative EVS, which can indicate overfitting. A RandomForest Regressor proved to be the highest performing underlying model with an average RMSE of 1088.16 and an EVS of 0.13829787. I chose to also use a RandomForest Regressor because it was simple to implement with the multi-step forecasting method as well as being "highly flexible" meaning it can avoid overfitting data unlike a logistic regression. I then used this method with a RandomForest Regressor to make a final submission to Kaggle.

I also considered some other options such as omitting data from China altogether because it does not reflect the growth rate of most other countries at the moment. However because of the required rows for Kaggle, this could not be done.

Conclusion

After submission to Kaggle I received a score of 2.31179. While this is not a very high score I was happy with being able to improve my predictions by implementing different methods and trying other underlying models. Kernel cache memory as well as computing power is shown to directly affect the performance of models, especially Support Vector Machines. Perhaps if I had more cache as well as computing power via the School of Computer Science lab machines I would have been able to improve performance even more. If given more data and a better way of multi-step forecasting I may have also been able to extract more important features.

References

3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.23.0 documentation. Scikit-learn.org. (2020). Retrieved 13 May 2020, from https://scikit-learn.org/stable/modules/model_evaluation.html#explained-variance-score.

COVID-19 death rate vs. Population density. Our World in Data. (2020). Retrieved 13 May 2020, from https://ourworldindata.org/grapher/covid-19-death-rate-vs-population-density.

Dowd, J., Rotondi, V., Andriano, L., Brazel, D., Block, P., & Ding, X. et al. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. https://doi.org/10.1101/2020.03.15.20036293

Drakos, G. (2019). *Random Forest Regressor explained in depth*. GDCoder. Retrieved 13 May 2020, from https://gdcoder.com/random-forest-regressor-explained-in-depth/.

Mahase, E. (2020). Covid-19: death rate is 0.66% and increases with age, study estimates. *BMJ*, m1327. https://doi.org/10.1136/bmj.m1327

Rocklöv, J., & Sjödin, H. (2020). High population densities catalyse the spread of COVID-19. *Journal Of Travel Medicine*. https://doi.org/10.1093/jtm/taaa038

Appendix: