How Countries' Macro-Data Inform Their Suicide Rates

Zeyan Du
Brown University

I. Introduction

Due to such factors as pace of development, technological advancement and economic growth, some countries seem to see a higher quality of life than others, at least per the macro-data. But do the people within countries with these better measures live more satisfying lives, as reflected by likelihood of committing suicide? In order to explore the answers, I looked into 101 countries' macro-data to find out how suicide rate is related to different features of the countries and its citizens. In doing so, I explored the signs that indicate increase in suicide rates both on a international and country-specific level.

The original dataset was compiled from four international organization datasets and contains 27,820 entries and twelve features where both numerical and categorical data are documented. After modifying the dataset, I kept seven features - country, year, sex, age, population, gdp_per_capita ($), and generation - and the target variable suicides/100k pop ("Suicide Rate"), which stands for the number of suicides for every hundred thousand people in each country, or in each sex or age group within each country. The problem is regression-based and it is important in that it tells us the features associated with high suicide rate so as to better inform what drives suicides from a macro-policy perspective, with the cumulative goal of preventing future suicides.

First I looked into how suicide rate changes with year to get the dynamics of suicide rate from 1985 to 2016. Then I examined the relationship between suicide rate and features to identify features strongly positvely correlated with suicide rate. Next, I evaluated countries with high suicide rates and compared their attributes against countries that have low suicide rates. In the end, I evaluated three regression models and used evaluation metrics to find the model that best fit the dataset.

Numerical Data:

- year: year ranging from 1985 to 2016
- population: number of people
- suicides/100k population: suicide rate for every 100k population (target variable)
- gdp_per_capita ($): GDP per capita

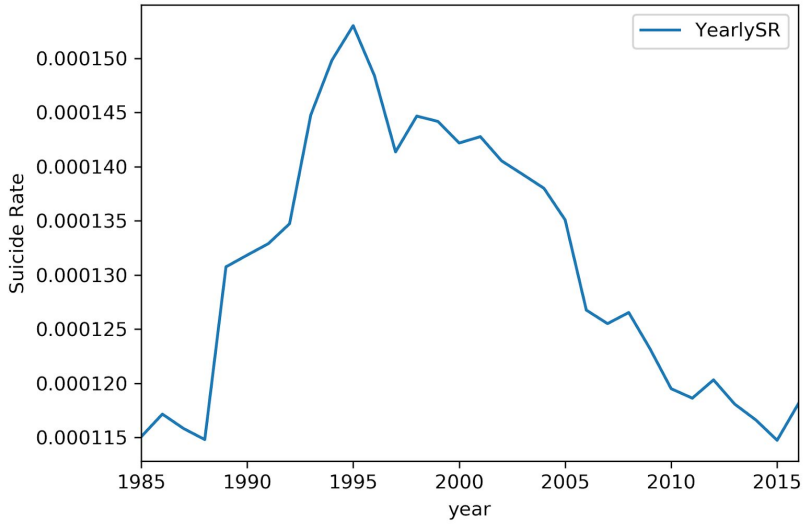
Categorical Data:

- country: 101 countries

- sex: male/female
- age: 6 age groups - 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, and 75+ years
- generation: 6 generations documented: G.I. Generation, Silent Generation, Boomers, Generation X, Millennials, and Generation Z

II. EDA

Suicide rate increased rapidly starting around 1988 up until around 1995, at which point it reached its peak. The ensuing drop lasted until 2015. A sign of an upward trend in suicide rate arises following 2015.



In order to get an overall perspective of the features most correlated with the target variable "Suicide Rate", I first created a correlation matrix excluding the 101 country variables. According to the correlation matrix (*Figure 1*), the top three features correlated with the target variable are "sex", "age" and "generation".

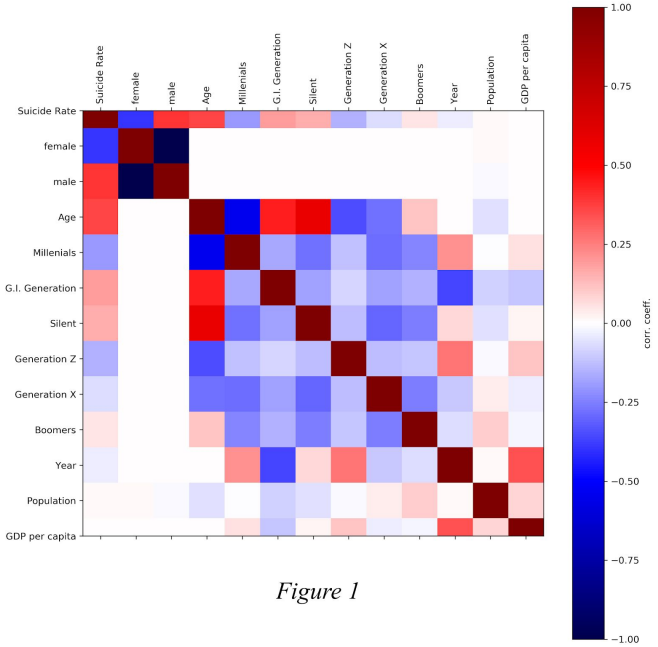


Figure 1

Suicide Rate vs Age

I created a pairs plot for a better visualization of the relationships between each feature (excluding the country variables) and target variable. According to the pairs plot (see *Figure 2*), the first correlation than stands out is "Suicide Rate vs Age". The graph shows that suicide rate is positively associated with age group. The table in *Figure 3* displays the mean suicide rate for each age group, which also confirms that suicide rate increases as age increases.

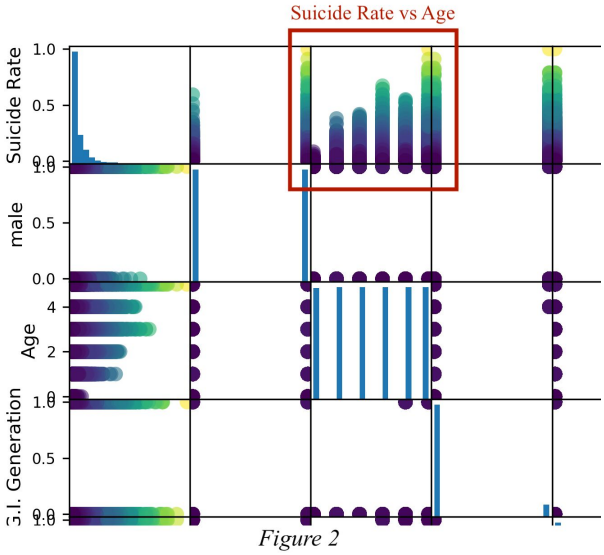


Figure 2

Age Group	Mean Suicide Rate
5-14 years	0.002756
15-24 years	0.039770
25-34 years	0.054171
35-54 years	0.066442
55-74 years	0.071812
75+ years	0.106483

Figure 3

Suicide Rate vs Generations

Thus, given these preliminary results, I decided to explore deeper into the suicide rate within each generation. There are six generations documented in total: G.I. Generation, Silent Generation, Boomers, Generation X, Millennials, and Generation Z. Based on the boxplot (see *Figure 4*) for each generation, the mean suicide rate decreases as the generation becomes more recent.

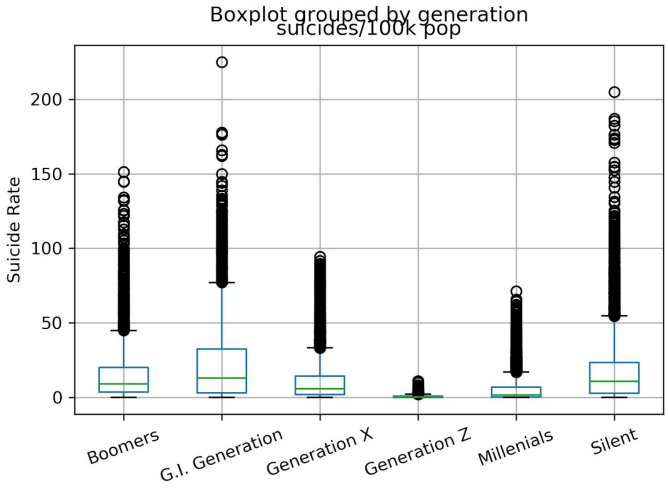


Figure 4

<u>Suicide Rate vs Sex</u>

Another noteworthy correlation from the pairs plot (see *Figure 5*) is "Suicide Rate vs Sex". Based on the boxplot (see *Figure 6*) for "Suicide vs Sex", it is clear that males are signficantly more likely to commit suicide than are females.

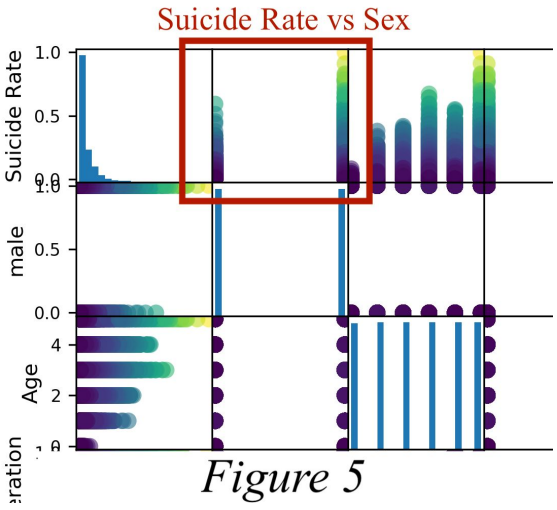


Figure 5

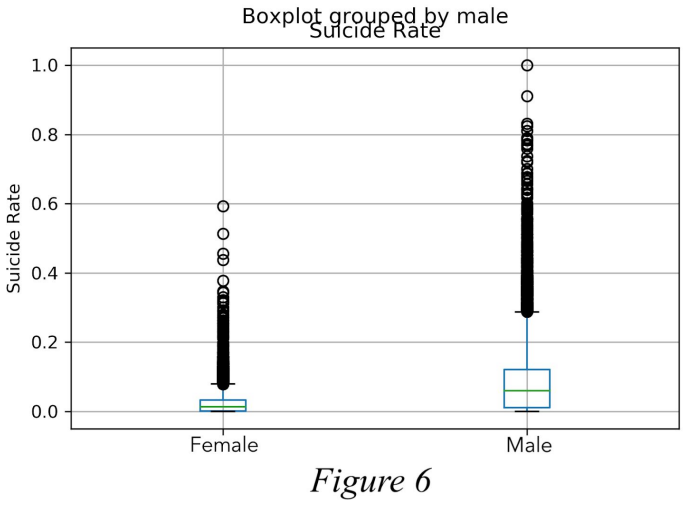
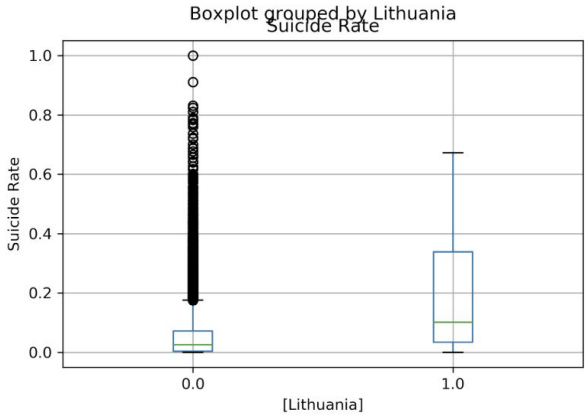
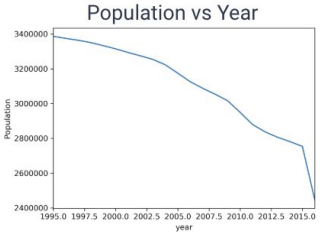


Figure 6

<u>Suicide Rate vs Countries</u>

With this information, I further explored how suicide rate differs among the various countries. The dataset consists of 101 countries. The f-test suggests that, among the countries within the dataset, the three countries with the highest suicide rates are Lithuania, Russian Federation, and Hungary. The country with the lowest suicide rate is Antigua and Barbuda. The boxplots and tables (see *Figure 7-10*) below record the comparisons of suicide rates between each of these countries and other countries. Another compelling finding I encountered upon by evaluating the change in population among these countries from 1958 to 2016 is that countries with higher suicide rates experience great decreases in population and countries with lower suicide rates have been experiencing constant increases in population.

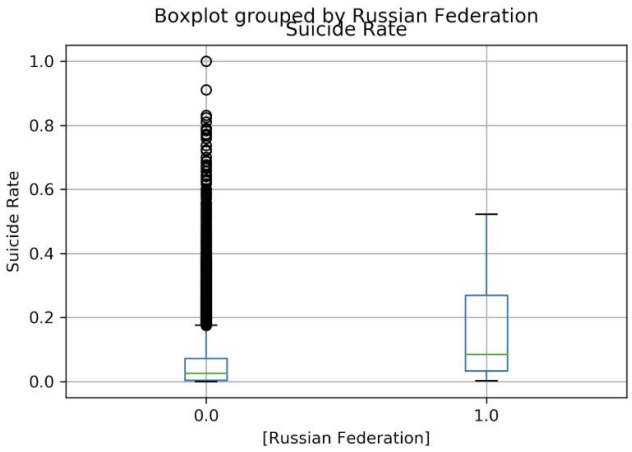


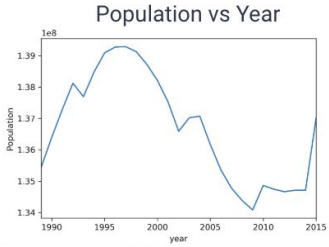


Lithuania	Mean Suicide Rate
0.0	0.055802
1.0	0.179649

68.94% higher

Figure 7 Suicide Rate in Lithuania

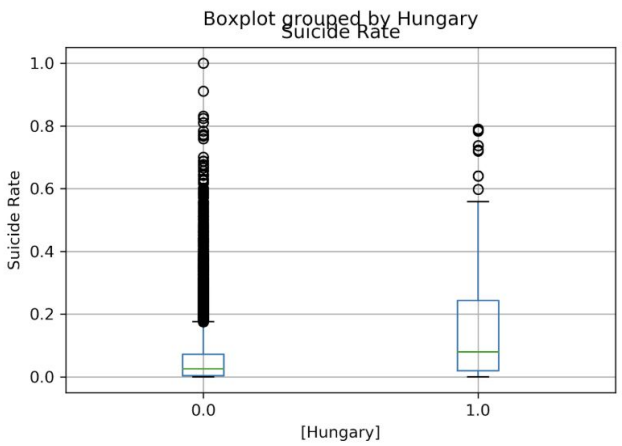


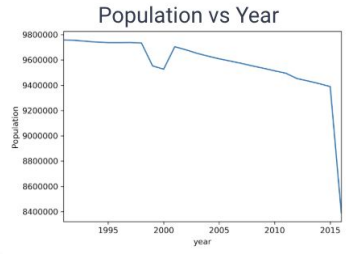


Russian Federation	Mean Suicide Rate
0.0	0.055812
1.0	0.155098

64.02% higher

Figure 8 Suicide Rate in Russian Federation

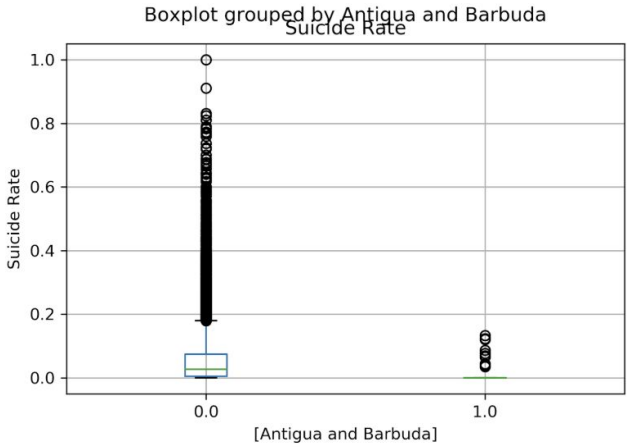


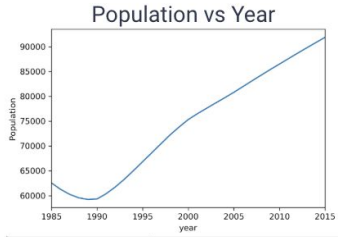


Hungary	Mean Suicide Rate
0.0	0.055969
1.0	0.145626

61.57% higher

Figure 9 Suicide Rate in Hungary





Antigua and Barbuda	Mean Suicide Rate
0.0	0.055097
1.0	0.065529

95.73% lower

Figure 10 Suicide Rate in Antigua and Barbuda

Suicide Rate vs GDP per capita

The above results call for an understanding of the causes behind the differences in suicide rates among these countries. One important factor that differs among the countries is GDP. Therefore I looked at the GDP per capita in these four countries and found out that before the year 2000, Lithuania and Hungary experienced slow increases in GDP per capita. Before 2000, Russian Federation saw decreases in its GDP per capita. Note that this time range is also where suicide rates increase the fastest. However, Antigua and Barbuda has gone through a rapid increase in GDP per capita during the same time. (see *Figure 11*)

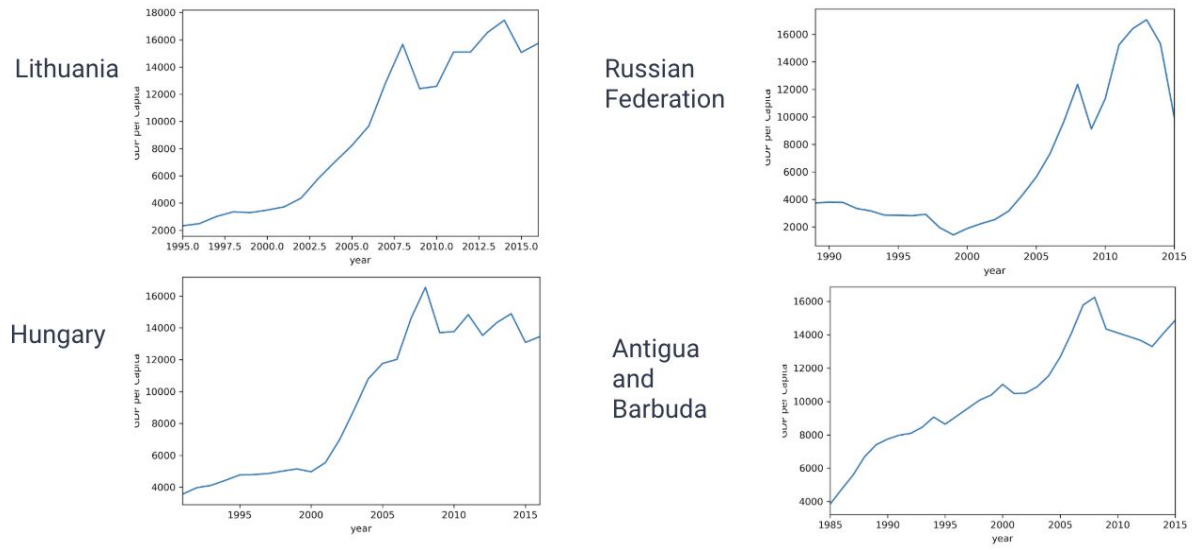


Figure 11

III. Methods

Data Preprocessing

After preprocessing the dataset, there are 27,280 features and 115 features. I used standard scalers for the features "population" and "gdp_per_capita ($)" because they are continuous variables and unbounded. One hot encoder was applied to the features "country", "sex" and "generation" due to their categorical and unordered attributes. I used ordinal encoder for the feature "age" because in this case, "age" represents six age groups so it is categorical and therefore can be ordered. Also, minmax scaler was applied for the feature "year" because the feature is continuous and bounded.

Numerical Data:

- population: normalized population
- gdp_per_capita: normalized GDP per capita of each country
- year: the values get larger as the years get more recent

Categorical Data:

- country: 101 columns with each column representing a country
- sex: 2 columns with one column representing male and another representing female
- generation: 6 columns, with each column representing each generation
- age: 6 age groups ranked from 0 to 5, with 0 representing the youngest age group and 5 representing the eldest age group.

ML pipeline

Since this is a regression problem, I considered two ML models: ridge regression, and random forest regression. The data is split as follows: 80% assigned to training data and 20% to test data. Then I used K-fold with five splits to evaluate which model best fit the dataset.

Ridge Regression - The parameter tuned for the ridge regression is α. I evaluated α within the range -16 (inclusive) to 1 (exclusive) with 100 numbers evenly spaced on a log scale to find the best α under 10 different random states.

Random Forest Regression - The parameters tuned for the random forest regression are max_depth and min_samples_split. Different values of max_depth and min_samples_split are evaluated under 10 different random states in the range [1,10) and [2,15) respectively.

The models I evaluated are regression models, so I used an evaluation metric of R2 score for both models to provide checks on how well the models fit the dataset.

Uncertainties

With respect to the uncertainties that may arise in a random forest regressor, I evaluated 10 different random states [0, 42, 84, 126, 168, 210, 252, 294, 336, 378] for each model and found the average and standard deviation of the test scores for each model.

IV. Results

R2 Score

Model	Test Score
Ridge Regression	0.509 +/- 0.009
Random Forest Regression	0.742 +/- 0.015

The baseline R2 score is 0. According to the table above, random forest regressor has a significantly higher R2 score than the ridge regression, indicating that random forest regression better fits the model.

Global Feature Importance

According to the graphs below (see *Figures 12*), the features are arranged in the order of importance.

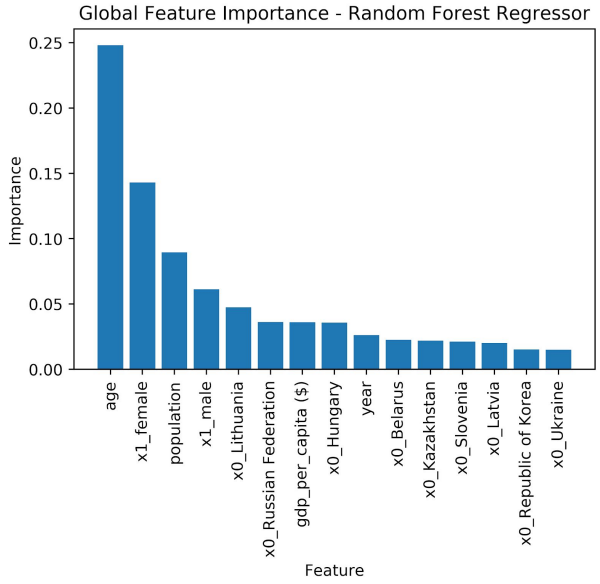
Age - Note that in all four graphs, age is the feature with highest importance, meaning that age has strong correlation with suicide rate. This also implies that suicide rate is strongly correlated with generations. According to the previous analysis, G. I generation and silent generation are the two generations with the highest suicide rates. These are the generations that experienced weighty global events such as World War II and the Great Depression, which caused major output decline, unemployment, and deflation in much of the world beyond the U.S.

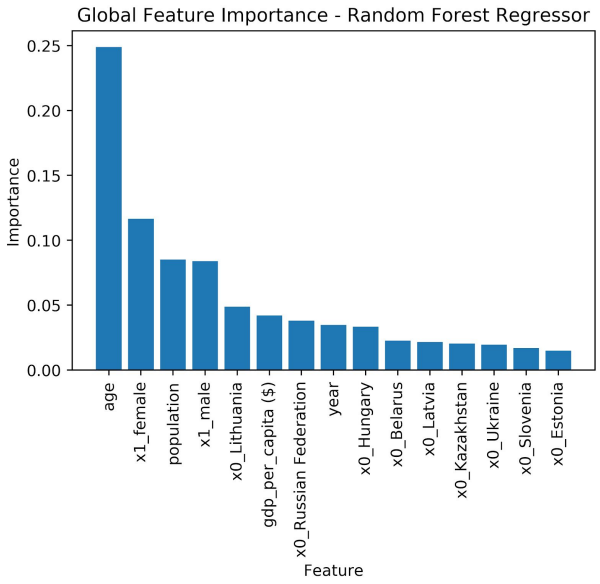
Sex - Sex came in second importance among all features. This proves the previous point made above that males are more likely to commit suicide than females. This phenomenon could be explained by 1) how globally speaking, males overall are under more pressure as breadwinners and soldiers, especially during periods of economic collapse or war; 2) personality and temperament differentials based on sex.

Population - Another factor strongly associated with suicide rate is population. In countries with a higher population, wealth inequality is greater and the stress of competition tends to be higher.

GDP per capita - GDP per capita is strongly associated with suicide rate. The average income/purchasing power of individual citizens, though does not necessarily indicate their overall happiness, is a useful indicator for standard of living and overall quality of life, which closely tie with happiness and personal well-being.

Countries - A few countries strongly correlated with suicide rate is shown in the feature importance graph. These are the countries that have high suicide rate.





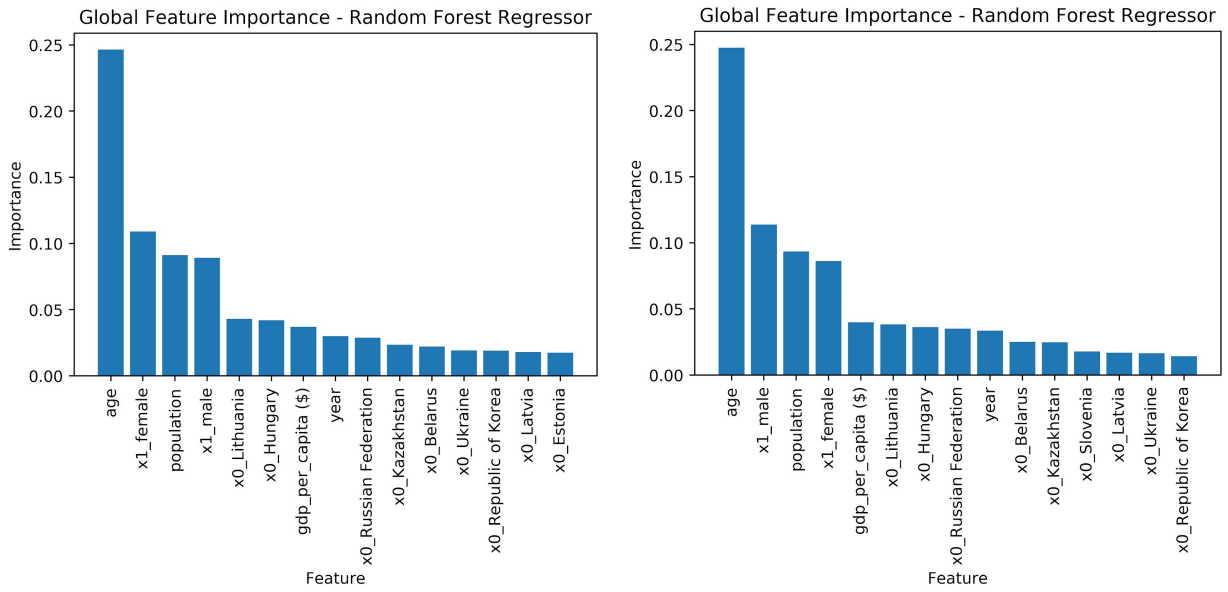


Figure 12

V. Outlook

In order to improve the model, I would figure out a better range to tune the parameters and go beyond the features associated with the 101 countries to evaluate how other features affect the target variable in different countries. Moreover, I would also consider interactions between the features to take into consideration the effects of correlations among features and see if any combinations of features have strong associations with the target variable.

VI. References

United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

[Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

The World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

Kaggle (Data Source). Retrived from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

VII. Github Repo

Github: https://github.com/daisydu97/DATA1030-Project.git