Statistical Analysis on Factors influencing Life Expectancy

Priyank Koul, Arif Sidiq Wani, Shevgoor Adithya Kamath, and Ashutosh Rout
PES University Bengaluru, priyankkoul14, wanixarif, kingdomofadithya, ashu24159@gmail.com

Abstract - **Initially in past, there has been a lot of study taken place on many factors influencing the mortality and the life expectancy, with their dependence to factors and variables which might be income-specific, or demographic. We realised that there were certain variables that were not being taken into consideration. Some of these variables would be vaccination/immunization and HDI (Human Development Index). Also, most of the previous research done had been performed by taking data of just a single year itself, for all the nations. So eventually, this research paper gives us the motivation to solve the issues of both the factors by formulating some regression models, by considering the data from a period of 2000 to 2015 for all the countries. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in letting countries understand which factors should be given more importance in order to efficiently improve the life expectancy of its population.**

INTRODUCTION

This is annual compilation of WHO's data. The series is produced by the WHO Department of Health Statistics and data Systems of the Health Systems and Innovation Cluster. As in previous years, World Health Statistics 2015 has been compiled using publications and databases produced and maintained by WHO technical programmes and regional offices. variety of demographic and socioeconomic statistics have also been derived from data produced and maintained by a variety of national and international organizations. The latter include the international organization Children's Fund (UNICEF), the international organization Department of Economic and Social Affairs (UNDESA) and its Population Division, the international organization Educational, Scientific and Cultural Organization (UNESCO), the global organization International Telecommunication Union (ITU) and the World Bank. the indications utilized in this report are included on the premise of their relevance to global public health, on data availability and quality, and on the reliability and comparability of the resulting estimates. Taken together, these indicators provide a comprehensive summary of the

present status of national health and health systems within the following nine areas:

- life expectancy and mortality
- cause-specific mortality and morbidity
- selected infectious diseases
- health service coverage
- risk factors
- health systems
- health expenditure
- health inequities
- demographic and socioeconomic statistics.

Since our dataset is published by WHO still, we are able to examine what they need drained order to work out some statistics, and that we can forestall for similar inferences, and a few more of them. they need summarised the statistics for various countries, and have done a comparison amongst veracious aspects. The report represents some best estimates of WHO for a broad range of key public health indicators supported evidence available at the start of 2015.

PREVIOUS WORK – A BRIEF REVIEW OF ONLY THE MOST RELEVANT PREDECESSOR WORK; WORLD HEALTH STATISTICS REPORT

For WHO, 2015 represents the target year for the Millennium Development Goals (MDGs). It mostly relates to factors such as Life Expectancy, Mortality Rates.[1]

MAIN CLAIMS

- The MDG on drinking-water called for the proportion of the population without sustainable access to safe drinking-water to be halved between 1990 and 2015, and was met globally in 2010.
- Progress in child survival worldwide has been described as one of the greatest success stories of international development, with child deaths being almost halved over the last two decades compared to the 1990 MDG baseline.
- Between 1990 and 2013 under-five death rates declined by 49%, tumbling from an expected 90 per 1000 live births to 46 per 1000 live births.[2]

MAIN TAKEAWAY

Most of the aspects and factors analysed in the dataset are similar to those of ours, considering the dataset was released by WHO too. So, whatever tables which are mentioned in the summary, we are planning to go with the same technique of analytics and infer things of some countries and then later comparing them.

REPRODUCTIVE AND CHILD HEALTH

This report was created to exhibit best practices in detailing the consequences of wellbeing disparity observing, and to present imaginative, intuitive ways for crowds to investigate imbalance information. Until recently, development goals and agendas have lacked a systematic focus on the reduction

[2] of within-country inequality. This report focuses on inequality in low- and middle-income countries.

MAIN CLAIMS

●An understanding of the state of inequality reveals gaps in population health and lends insight into how policies, programmes and practices can be aligned to promote the ideal of health for all.

●When the data is broken down by subgroups differences between social groups, that might have otherwise remained hidden behind the overall average, are revealed.

MAIN TAKEAWAY

With this it would be possible to analyse to infer if certain goals and targets which were planned were quite achieved or not.

TRENDS IN OLDER CHILDREN AND YOUNG ADOLESCENTS: AN ANALYSIS [3]

Little is known on how the mortality risk among older children and young adolescents has changed from 1990 to 2016. This report estimated trends in mortality of older children and young adolescents in the 195 countries from 1990 to 2016.

MAIN CLAIMS

●The paper claims that the top five global causes of death in children aged 5–14 years in 2015 were lower respiratory tract infections, diarrhoeal diseases, drowning, meningitis, and road injuries. This indicates that with public health interventions substantial progress can be pulled of for this age group.

● Since 2000, progress in reducing mortality among older children and young adolescents has not been in focus when comparing with children younger than 5 years.

● The risk of mortality of children aged 5–14 years has substantially decreased since 1990, despite this age range not being specifically targeted by health interventions.

MAIN TAKEAWAY

The paper claims that in low-income and middle-income countries, mortality of older children and adolescents comprises about 98% of all mortality of older children and adolescents in 195 countries, which comprise 89.5% of the global population of children of this age group. [4]

TRENDS IN NEONATAL MORTALITY

An essential part of the third Sustainable Development Goal (SDG) is to end preventable child deaths. To achieve this goal the paper believes that an understanding of the trends in neonatal mortality.

MAIN CLAIMS

● Estimate the trends in neonatal mortality with a statistical model to assess progress in the SDG era. With these estimates the paper then assesses different variables that would affect NMR for the period 2018 to 2030.

MAIN TAKEAWAY

The paper believes improvements are needed in the countries with high NMR, specifically parts of Sub-Saharan Africa and South Asia.

WHAT IS THE PROBLEM AREA?

It is quite obvious that all the global health organisations, have been looking forward to a common goal, which would be to increase the living conditions, mostly in terms of health aspects, and eventually provide healthy and disease-free surroundings for majority of the world population. Also, it is quite evident that if people are provided with such facilities, the life expectancy is definitely going to increase.

So, we are here planning to do some analysis on Life Expectancies, and try to predict how accurately are we predicting the life expectancies, and which features in our dataset are contributing to what extent.

INITIAL STEPS

Initially, for convenience, we decided to manipulate our dataset to an extent so that it later becomes easier for us to work upon. We changed our column names to all lower case for simplicity, and made all of them accessible in an easier way.

We also figured out that there are quite a lot of outliers in our data, and these might affect the accuracy or domain specificity of our model to great extent. We plotted some box plots to visually analyse some of the column's outliers. We would soon try to handle them.

```
country                          0
year                             0
status                           0
life_expectancy                 10
adult_mortality                 10
infant_deaths                    0
alcohol                        194
percentage_expenditure           0
hepatitis_b                    553
measles                          0
bmi                             34
under-five_deaths                0
polio                           19
total_expenditure              226
diphtheria                      19
hiv/aids                         0
gdp                            448
population                     652
thinness_10-19_years            34
thinness_5-9_years              34
income_composition_of_resources 167
schooling                      163
dtype: int64
```

These are the number of null values we are seeing in each column individually. Let's see how we deal with them. But before that, we ought to apply some domain knowledge to the dataset and realise that there are many other

redundant features, which might put up a slight inaccuracy in the dataset itself. So, we saw the details, and decided to replace some of the 0 values in the column with NaN, and in BMI column, we are replacing the values with NaN if the value is not in a certain threshold we have provided.
After checking that, we realised that the number of NaN values in BMI column was more than 50%, hence we decided to drop the column itself.
Now, to handle out rest of the missing values, we would impute those places with the mean of the column value, but with respect to that particular year itself. Because, obviously that would be a better option to consider in our case.
After we have dealt with the columns and their missing values, the point comes here to deal with the outliers. We run some code to see what are the percentages of the outliers. Following are the results we observed:

```
---------------life_expectancy---------------
Number of outliers: 17
Percent of data that is outlier: 0.58%
---------------adult_mortality---------------
Number of outliers: 97
Percent of data that is outlier: 3.3%
---------------infant_deaths---------------
Number of outliers: 135
Percent of data that is outlier: 4.59%
---------------alcohol---------------
Number of outliers: 3
Percent of data that is outlier: 0.1%
---------------percentage_expenditure----------
Number of outliers: 389
Percent of data that is outlier: 13.24%
---------------hepatitis_b---------------
Number of outliers: 222
Percent of data that is outlier: 7.56%
---------------measles---------------
Number of outliers: 542
Percent of data that is outlier: 18.45%
---------------under-five_deaths---------------
Number of outliers: 142
Percent of data that is outlier: 4.83%
---------------polio---------------
Number of outliers: 279
Percent of data that is outlier: 9.5%
---------------total_expenditure---------------
Number of outliers: 51
Percent of data that is outlier: 1.74%
---------------diphtheria---------------
Number of outliers: 298
Percent of data that is outlier: 10.14%
---------------hiv/aids---------------
Number of outliers: 542
Percent of data that is outlier: 18.45%
---------------gdp---------------
Number of outliers: 300
Percent of data that is outlier: 10.21%
---------------population---------------
Number of outliers: 203
Percent of data that is outlier: 6.91%
---------------thinness_10-19_years------------
Number of outliers: 100
Percent of data that is outlier: 3.4%
---------------thinness_5-9_years--------------
Number of outliers: 99
Percent of data that is outlier: 3.37%
---------------income_composition_of_resources-
Number of outliers: 130
Percent of data that is outlier: 4.42%
---------------schooling---------------
Number of outliers: 77
Percent of data that is outlier: 2.62%
```

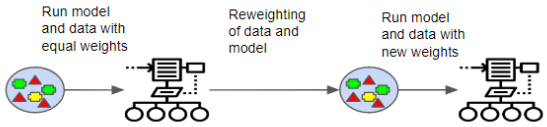
Now, we are on track and ready with what we will be trying to achieve. We have our dataset ready.

FURTHER STEPS

First, we would remove our categorical columns. So, we drop the Country and the Status columns from our data frame. Now, we are going to create our X variables and Y variables. X variables as we know are the independent variables as we have studied, and Y is the target variable, or the dependent variable. Since we already mentioned that we would be working to predict Life Expectancy, so every other column in our dataset would be the independent variables except for the Life Expectancy, and Life Expectancy itself is going to be the target variable. Now, for training any ML model, when it is supervised, after we are done training, we would need to have some new dataset, with which we can test our values and come to certain conclusion if our model has been performing well. We would use Python's scikit learn module, to divide our existing dataset into a training and a test dataset itself. In our code, we are taking 20% of the dataset as the test dataset, and the rest 80% is being used as the training dataset.

GRADIENT BOOSTED REGRESSOR

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
How Gradient Boosting basically works is by Gradient Boosting builds an additive model in a forward stage-wise manner, as it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.



So, if we kind of analyse and see how Gradient Boosting works, initially we select a learner. Most likely, the learner is going to be weak. Now, fundamentally how we take a combination of models, and then later use an additive model to boost the working of our model. To check the accuracy, we obviously would need to define a loss function, and further like we do in all the ML models, we would end up in minimising that loss function itself.
There are some advantages of Gradient Boosting, such as better accuracy (of course), comparatively lesser need for pre-processing, and a great extent of higher flexibility. Also, missing data does not pose any problem to GBR as the algorithm knows how to deal with it well.
So, now back to our research, we import certain loss functions such as the mean absolute error, and the mean squared error, which we use to minimise the errors of our Gradient Boosting Algorithm, further r2 score is a metric function which was imported. It is the proportion of the

variance in the dependent variable that is predictable from the independent variable.
We would use the fit () function, to fit the model with respect to the training X and training Y we had set initially. When it is done, we use predict () to predict the Y values of our test dataset. Once done, we check it with the actual Y values of the dataset, and using the loss functions, we compute the r2 score.
Our initial r2 comes out to be 0.94, which is quite decent, but now since it is a research-based project, we would be looking forward to improve the accuracy.
In the previous part of the project, we had mentioned there were certain outliers in some columns, and I had mentioned at a later point of time, we would have to take care of them. Now is the time. We follow the conventional theory that anything that does not lie in the range of first quartile minus the 1.5 times the Inter Quartile Range, and last quartile plus the 1.5 times the Inter Quartile Range is an outlier.

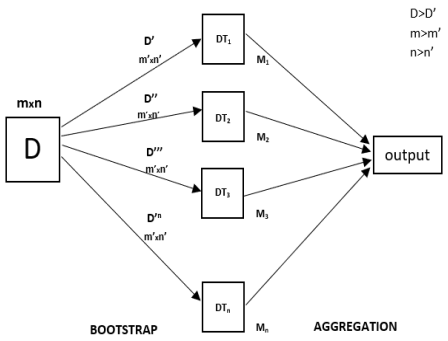
```
(                                  0
 adult_mortality                  97
 alcohol                           3
 country                           0
 diphtheria                      298
 gdp                             300
 hepatitis_b                     222
 hiv/aids                        542
 income_composition_of_resources  130
 infant_deaths                   135
 life_expectancy                  17
 measles                         542
 percentage_expenditure          389
 polio                           279
 population                      203
 schooling                        77
 status                            0
 thinness_10-19_years            100
 thinness_5-9_years               99
 total_expenditure                51
 under-five_deaths               142
 year                              0,
 Columns: [0]
 Index: [], Empty DataFrame
 Columns: [0]
 Index: [], Empty DataFrame
 Columns: [0]
 Index: [])
```

Here, surprisingly or not, almost all the columns have outliers, and obviously we wouldn't want to mess up with the constituents of our dataset. So, here we would end up making a decision that instead of taking care of the outliers, we would be using Random Forest Regressor, since they are very robust to outliers.
There are two pieces of intuition we would prefer knowing in the first place:

1.) Whenever we a construct a decision tree, we must classify all the points. What is means that even outliers will get classified, and most likely would affect the decision trees at a point they were chosen during the boosting proces.
2.) A part of how a RandomForest does sub-sampling is called Bootstrapping, which is quite robust to outliers.

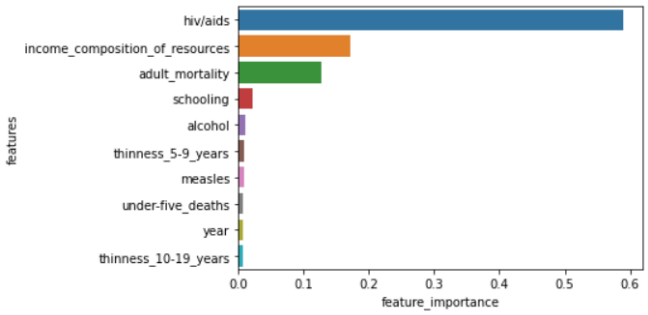
RANDOM FOREST REGRESSOR

So, basically it is the base learner, which is robust to outliers. They're the decision trees in our case. They isolate some of the very different observations into smaller leaves, which means smaller sub spaces of the actual space. In a regression problem, it is mostly a low order regression model, and for classification, it is voting.
Hence, for regression, extreme values/outliers are not affected by the entire model because they get averaged locally. Thus, technically the fit to the other values is not affected.



In our case, we import Random Forest Regressor, and then we fit our training and test sets and it eventually creates the model.
Later on, we can predict the values, and see how much they deviate from the actual values. And in this as well, we go on to calculate the R2 score as the metric.
For the Random Forest Regressor, the R2 score comes out to be 0.96, which is better than our Gradient Booster. This proves the fact that this turned out to be more robust to Outliers.

Further, we do not just stop here.
We look forward to observing the feature importance. Feature importance is very important when it comes to huge data, because a lot of features which might just be redundant, would just increase the in-memory computations, and would just be an overhead for us.
In our case, we took the top 10 features to see.
We plotted the features with their importance, and this is what we inferred:

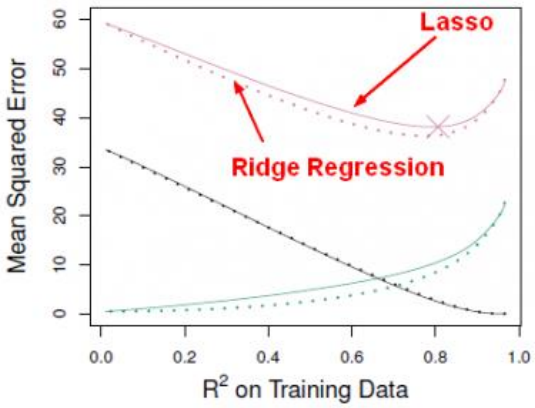


Here, as we can see, these are the top 10 features with the feature importance. Now, for a simple experiment, we repeated our same Random Forest Regressor model, by just taking these 10 features and s how well it performed.
It turned out by reducing the number of features to 10, we still got the same R2 score, in our case being 0.964.

MULTIPLE LINEAR REGRESSION

Multiple Linear Regression is a supervised way of training our dataset and then predicting our target feature on the test dataset.
We performed multiple linear regression, and found out the mean-squared error to be 2.09.
The R2 score came out to be 0.94, which again seemed to be a decent score.
However, here we can make a small inference that the R2 score here is little lesser than from the RF Regressor because of the presence of outliers.

RIDGE AND LASSO REGRESSION

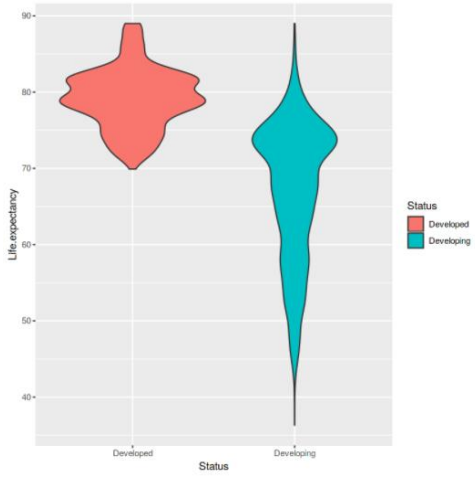


As a simple experiment, we tried out with Ridge and Lasso Regression as well, and both of them seemed to be giving off a decent R2 score as well, 0.94 in both cases.

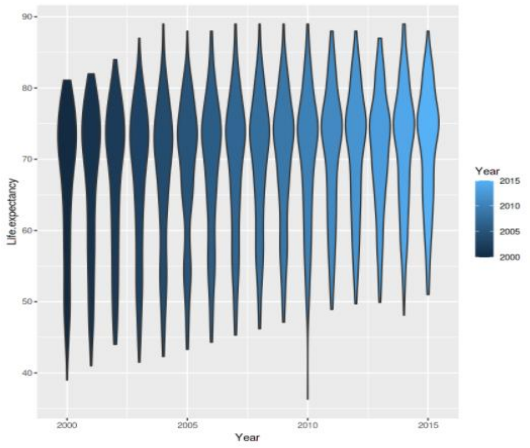
CONCLUSION

In this report, we analysed the WHO Life Expectancy dataset. We started off with handling and pre processing of data, and some basic EDA and visualisations in R, whose details have been mentioned in the initial report. Followed by we checked outliers, certain correlations, and tried to choose a good model which would work well on our dataset. We tried for Gradient Boosters, but since it couldn't handle outliers well, we went on to go for Random Forest Regressors, and it gave a slightly better accuracy because we found out that it is robust to outliers. Lastly, we went to see some feature importances and tried the same model for a reduced dataset with lesser features. We found that to have almost the same accuracy as before. And finally, we wrapped up our project research by performing Multiple Linear Regression, followed by Ridge/Lasso's regression. Priyank, being the main leader, came up with the idea of the dataset, and the ideas of research we could dive into during the entire project course. Arif and Ashutosh came up with visualisation techniques, and finding out certain correlations between different features. Adithya and Priyank figured out what sort of models to implement upon. All the team mates performed the literature survey on their own. Last but not the least, each of us worked together in a team, and we are highly thankful to our esteemed professors, without whom we definitely wouldn't be able to complete this.

APPENDIX



Status Division of Countries



Year -wise comparison of Life Expectancies

REFERENCES

[1] WHO World Health Statistics Report 2015.
[2] Global, regional, and national mortality trends in older children and young adolescents (5–14 years) from 1990 to 2016: an analysis of empirical data.: United Nations Inter-Agency Group for Child Mortality Estimation
[3] National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis.
[4] State of inequality: Reproductive, maternal, new born and child health. Published: 2015

BIBLIOGRAPHY

Danzon,Patricia M,and Jonathan D Ketcham,2003,Reference pricing of pharmaceuticals for Medicare.
Gage,Timothy B,,and Kathleen O'Connor.1994.Nutrition and the variation in level and age patterns of mortality.Human Biology 66:77-103.
House, J.S., R.C. Kessler, A.R. Herzog, R.P. Mero, A.M. Kinney, and
M.J. Breslow. 1992. "Social Stratification, Age, and Health." In K.W.
Schaie, D. Blazer, and J.S. House (Eds.), Aging, Health Behaviors, and
Health Outcomes. Hillsdale, NJ: Lawrence Erlbaum Associates.
Kovar, M.G., J.E. Fitti, and M.M. Chyba. 1992. "The Longitudinal Study
of Aging: 1984-90." Vital and Health Statistics, Series 1, No. 28.
Hyattsville, MD: National Center for Health Statistics