Factors influencing life expectancy globally

UHSE Machine learning boot camp 2020
(Garima Singh, Erika Andrade, Tony Tran, Mark Martir- Irizarry)

Problem Introduction

Life expectancy is a key parameter that reflects the well-being of people. There are several factors such as socio-economic, lifestyle habits, healthcare facilities which impact the human health.
The ultimate goal of the analysis is to predict 'Life expectancy' and suggest factors that are highly correlated to it and should be addressed by countries as well as at individual levels to achieve better life expectancy.

The main goals of this study are:

(1) Predict average life expectancy using the given features,
(2) Study the correlation of these features with life expectancy and with themselves.
(3) Study trends of 'Life expectancy' over the years from 2000 to 2015
(4) Understand what has affected life expectancy positively or negatively.
(5) Impact of the 'Status' of a country being a developed or developing have on 'Life expectancy?

Dataset

The dataset used for this project was obtained from Kaggle.com [1]. The data set consists of Life expectancy of 193 countries and related health and economic factors spanning over 15 years from 2000 to 2015. Hence, the dimension of the dataset is 2938x22. The original source of life expectancy and health data is the Global Health Organization (GHO) repository maintained by the World Health organization (WHO) and corresponding economic data was obtained from United Nations (UN).

Features and Processing

The list of features with Life expectancy (in age) is related are as below:

1) **Adult Mortality** Rates of both sexes (between 15 and 60 years per 1000 population)
2) Number of **Infant Deaths** per 1000 population
3) **Alcohol**, recorded per capita (15+) consumption (in litres of pure alcohol)
4) **Percentage Expenditure** on health as a percentage of Gross Domestic Product per capita(%)
5) **Hepatitis B** immunization coverage among 1-year-olds (%)
6) **Measles** - number of reported cases per 1000 population
7) **Body Mass Index** of entire population (percentage obese)
8) Number of **Under-five deaths** per 1000 population
9) **Polio** immunization coverage among 1-year-olds (%)
10) Government **expenditure percentage** on health of total government expenditure (%)
11) **Diphtheria** tetanus toxoid and pertussis immunization coverage among 1-year-olds (%)
12) Deaths per 1 000 live births **HIV/AIDS** (0-4 years)
13) Gross Domestic Product (**GDP**) per capita (in USD)
14) **Population** of the country
15) **Thinness 10-19 years**: Prevalence of thinness among Age 10 to 19 (%)
16) **Thinness 5-9 years**: Prevalence of thinness among Age 10 to 19 (%)
17) Human Development Index in terms of **income composition of resources** (index from 0 to 1)
18) Number of years of **Schooling**(years)

The data set has 2563 missing values and 535 outliers. GPD, population and Hepatitis B immunization alone accounted for 65% of the missing data. In lack of domain expertise, we have identified the outliers based on theoretical possible ranges of the features by their definition. We also tried the Interquartile Rule to eliminate the outliers, but it resulted in loss of two thirds of the data. Hence, we proceeded with the first approach. After identifying the missing values and outlier we followed two different approaches. **First approach:** Eliminate all the rows which had any missing values or an outlier. This strategy removed 54% of the data. **Second approach:** Eliminate row containing outliers and fill the missing values of a feature with country-wise mean values of all years. For some countries the feature value for all the years were missing, therefore they still had null values, such rows were removed from the dataset. This strategy removed 46% data.Therefore, the second approach saves us ~8% data.

Also, we plotted the dataset histograms to see the feature distribution. Some had normal distribution, BMI had bimodal distribution, while some had skewed data distribution.

In the next step of data processing, we normalized the dataset and determined the feature correlation matrix. Figure 1 shows the correlation of the factors with Life expectancy. Eleven factors are seven positive contributors to 'Life Expectancy'.

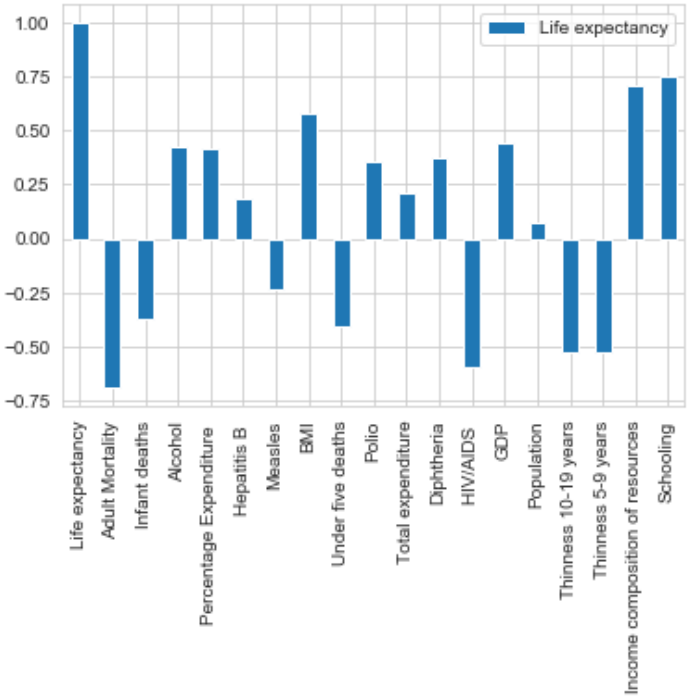

Figure 1 Correlation of the factors with Life expectancy

The features' correlation heatmap is as shown in figure below. It is interesting to note that, there are three pairs of features which are highly correlated to each other. There are: 'Under five deaths' and 'Infant deaths' per 1000 population (corr: 0.99); 'Percentage expenditure' of GDP per capita on health by government and 'GDP per capita' (corr: 0.97); Percentage 'Thinness prevalence 10-19' age and ' Thinness prevalence 5-19' age (corr: 0.97).

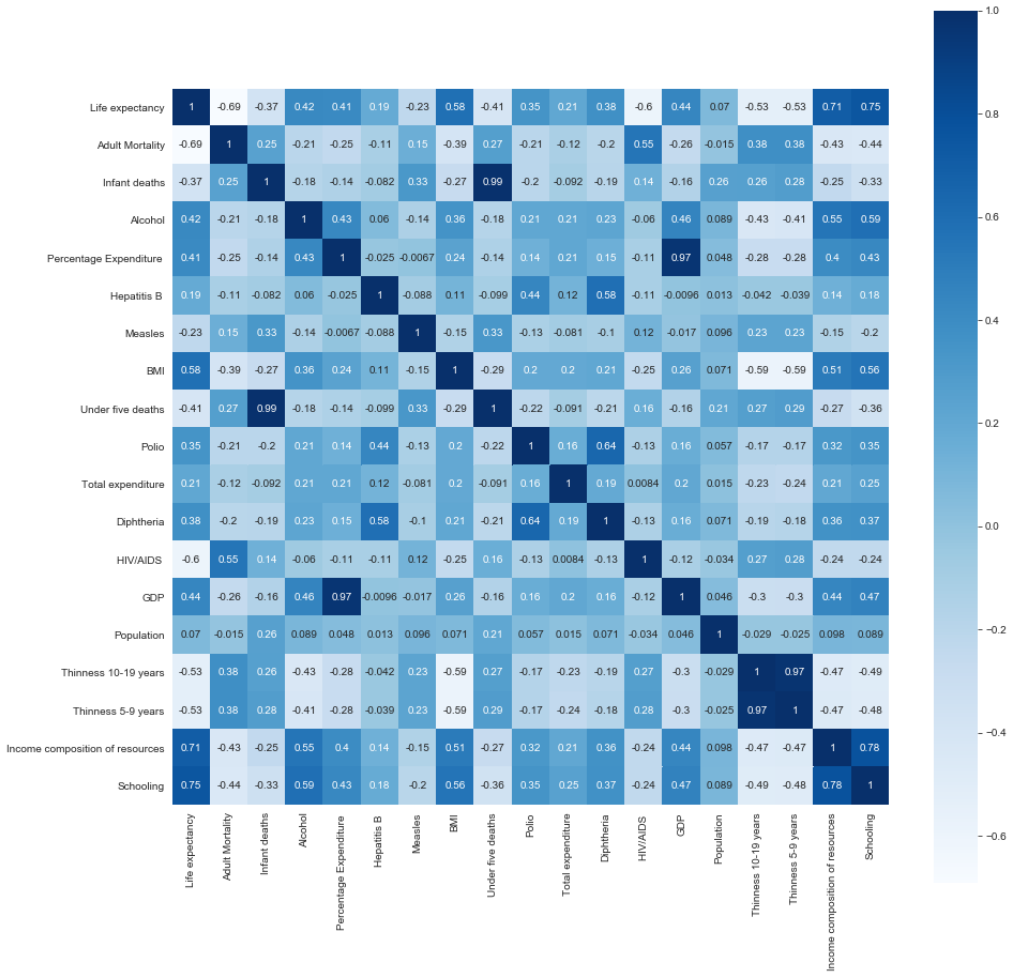

Figure 2 Correlation heatmap of all the features

Models and techniques

We split the data set into training set and training set in the ratio 80: 20 and use Multiple linear regression to predict the Life Expectancy. We further dropped one of the features of the correlated feature pairs and see if it improves or deteriorates the prediction accuracy.

Results and Discussion

There has been an increasing trend of Life expectancy over the years but there has been huge difference between the developed countries. Base on the correlation matrix, the countries can focus on the strongest impacting factors such as schooling years, immunizations, adult, infant deaths.

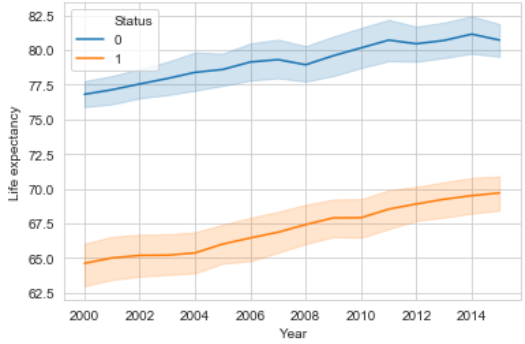

Figure 3 Global Life expectancy trend over last 15 years . Status 0 is 'Developed' and Status 1 is 'Developing'

The predictions of Life expectancy using different approaches of data cleaning are shown in and Figure 5 and Figure 5

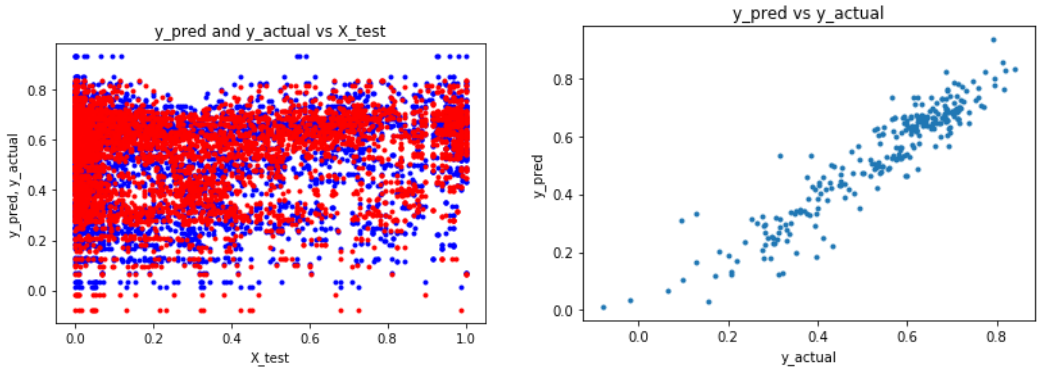

Figure 4 Prediction of Life expectancy (y) using dataset from with all the outliers and missing value rows removed. R_2 error = 0.85

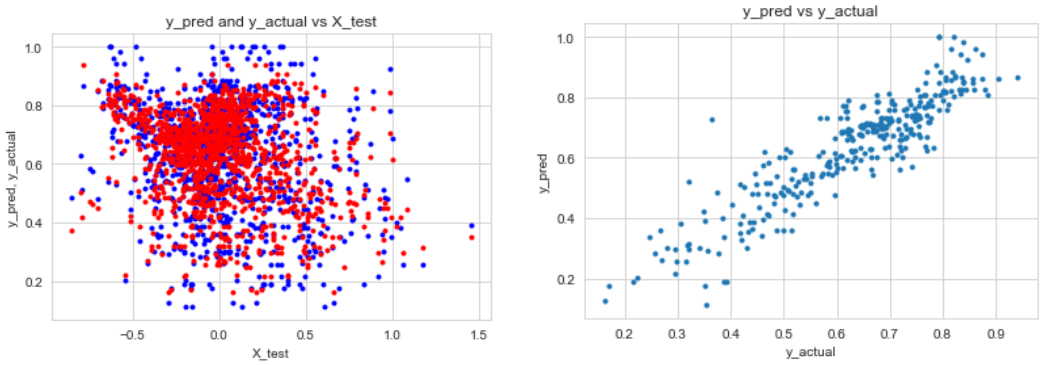

Figure 5 Prediction of Life expectancy (y) using dataset from with all the outliers rows removed and missing values filled with country mean. R_2 error = 0.77

Further, we dropped out one of the highly correlated pairs to study the effect on the prediction and found that only dropping out 'income composition of resources' had a significant detrimental effect on the predictions while others had negligible impact on it.

Future Work

In future, we would like to study the following:

1) Filling the missing values using KNN imputer
2) Use Standard feature scaling (using standard deviation and mean) as some of the features do not have a normal distribution
3) Dimension reduction using Principal component analysis
4) Use *f-test* to predict the model accuracy
5) Use non-linear regression methods to improve the model prediction