Peer-graded Assignment: Prediction of Life Expectancy

Introduction and main objective of analysis

For this course project a free data set from Kaggle website (available **here**) was chosen. The source of this data set is the Global Health Observatory (GHO) data repository under World Health Organization (WHO) and is used for the statistical analysis on factors influencing life expectancy, but in this assignment my main objective will be **prediction** of life expectancy based on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. It is also usefull to mention, what exactly does life expectancy mean:

Life expectancy at a given age represents the average number of years of life remaining if a group of persons at that age were to experience the mortality rates for a particular year over the course of their remaining life. Life expectancy at birth is a summary measure of the age- specific all cause mortality rates in an area in a given period.

In this analysis life expectancy at birth will be used as an outcome variable.

Description of the data set

The data were collected during the years 2000-2015 for 193 different countries and contains 2938 observations and 22 features. All predicting variables could be divided into 4 broad categories: Immunization related factors, Mortality factors, Economical and Social factors:

1. Immunization:
    - Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
    - Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)
    - Diphtheria - diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
    - HIV/AIDS - deaths per 1 000 live births HIV/AIDS (0-4 years)
    - Measles - number of reported measles cases per 1000 population
2. Mortality factors:

- Adult mortality - adult mortality rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- Infant deaths - number of infant deaths per 1000 population
- Under-five deaths - number of under-five deaths per 1000 population

3. Economical factors:
   - Status of country - developing or developed
   - GDP - Gross Domestic Product per capita (in USD)
   - Population - population of the country
   - Percentage expenditure - expenditure on health as a percentage of Gross Domestic Product per capita(%)
   - Total expenditure - general government expenditure on health as a percentage of total government expenditure (%)
4. Social factors:
   - Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
   - Schooling - number of years of Schooling (years)
   - Alcohol - alcohol consumption (in litres of pure alcohol) recorded per capita (15+)
   - BMI - Average Body Mass Index of entire population
   - thinness 1-19 years - prevalence of thinness among children and adolescents for Age 10 to 19 (%)
   - thinness 5-9 years - prevalence of thinness among children for Age 5 to 9(%)
5. Other: Country, Year
6. Outcome variable: Life expectancy at birth

Summary of data cleaning and exploration

As the data-sets were from WHO, no evident errors were found. On the other hand, a lot of data was missing. For features like Population, Hepatitis B and GDP around 500 values(!) were missing for less known countries: Vanuatu, Tonga, Togo, Cabo Verde etc. As a result it was decided to exclude these features from the final model. Also Total expenditure was excluded, because variable Percentage expenditure was used instead. The rest of missing values were simply deleted from the data set.

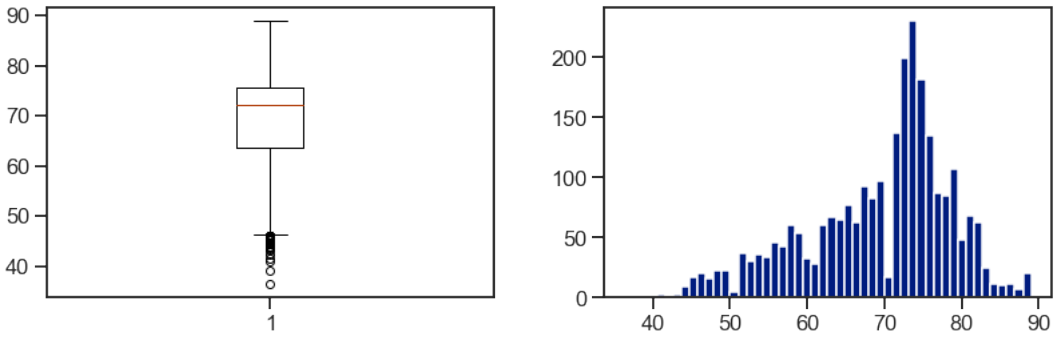
On the figure 1 box plot and histogram of Life expectancy are depicted. According to the box plot, there are 41 outliers - values lower than lower whisker (46.2 years). Such extremely low Life expectancy values were observed in developing countries like Sierra Leone, Swaziland, Malawi, Haiti etc. It was decided to remove them from the data set, because replacing outliers with median/mean value is not suitable for the outcome variable. Finally, the variable status of country was one-hot encoded.

Table 1: Missing values

Feature	Missing values
Adult Mortality	10
Polio	19
Diphtheria	19
thinness 5-9 years	34
thinness 1-19 years	34
BMI	34
Schooling	163
Income composition of resources	167
Alcohol	194
Total expenditure	226
GDP	448
Hepatitis B	553
Population	652

Figure 1: Outcome variable - Life expectancy

Box plot *Histogram*

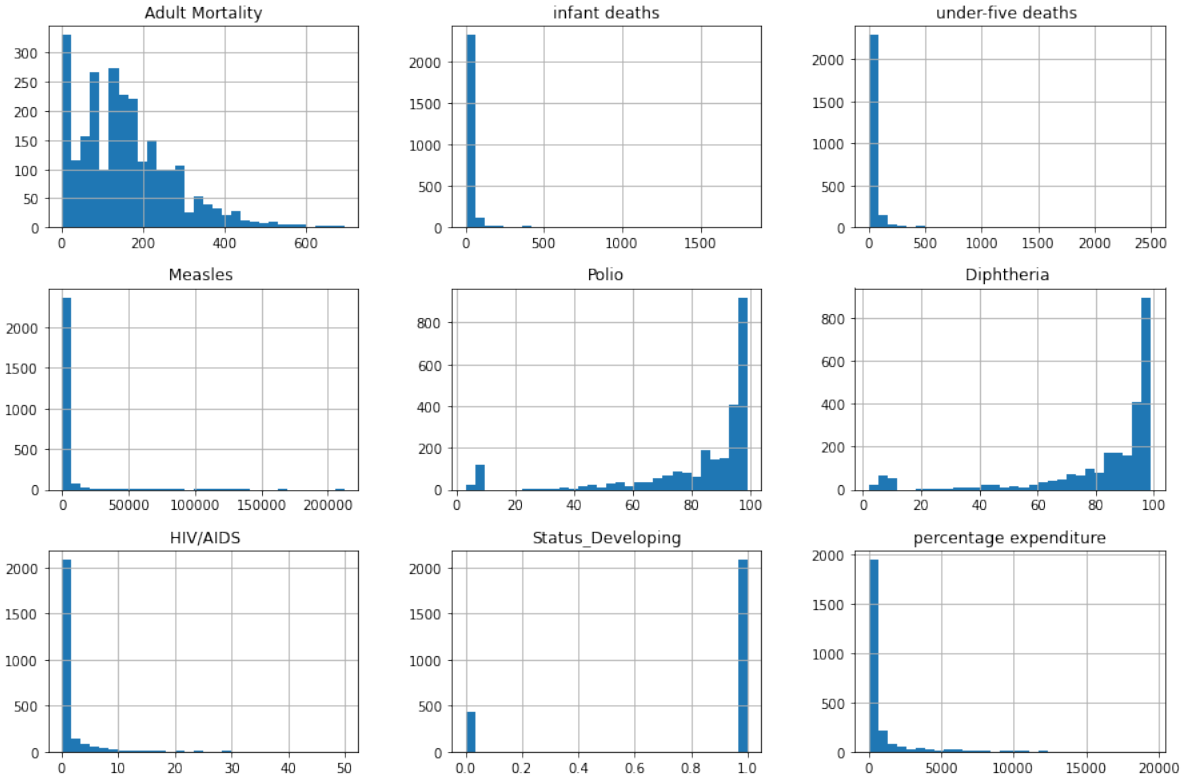


Moreover, the distribution of Life expectancy is clearly not normal and negatively skewed (see histogram in figure 1). I have tried to transform the dependent variable in order to normalize the distribution, but it wasn't successful. I have tried:

- Square-root transformation: sqrt(max(y+1) - y)
- Logarithmic transformation: log10(max(y+1) - y)
- Inverse: 1/(max(y+1) - y)

In the figures 2 and 3 below the histograms of main features are depicted. It can be seen clearly, that the prevalence of infant deaths, under 5 deaths and HIV/AIDS - deaths per 1 000 live births is almost 0 in majority of countries but reaches very high values for some developing countries. The same effect holds either for decease immunization (Polio, Diphtheria) and the number of Measles cases.

Figure 2: Histograms of main features



Key findings from training 4 linear regression models

After the exploratory data analysis and data cleaning, 4 linear regression models were fitted: simple linear regression as a baseline, polynomial regression and finally Ridge and Lasso regularisation regressions. 5-fold cross validation method was used in order to estimate how accurately the predictive models will perform in practice. Cross-validation averages measures of fitness in prediction and derives a more accurate estimate of model prediction performance. Also standard scaling was used for features data set in all 4 models. Features scaling does not have an impact on linear regression but does improve predictive performance in case of regularisation regression models.

I decided not to include Country into regression models, because there are 193 (!) countries in total and only 16 or less observations for each of them, which is not enough to estimate unbiased parameter for every country. Moreover, I used only degree 2 polynomial features for the linear regression, because in case of degree 3, there will be 969 coefficients estimated using only 2521 rows of data, and it is not enough. On the other hand, in case of regularization regression models like Lasso and Ridge, 3rd degree polynomial features could be used, because these method will zero out multiple coefficients (Lasso) or at least make them close to zero (Ridge). GridSearchCV together with Pipeline were used for estimation and iterating through different alpha values and polynomial features degrees.

Table 2 presents the results of different models estimation. The simple regression model has only 17 parameters and the smallest R squared value (83%), which means that this model explains 83% of total variability of Life expectancy. Polynomial model has significantly higher R squared (91%), and finally Ridge and lasso models have the highest R squared values. Hence I will chose Lasso regression as the final model, because it has

Figure 3: Histograms of main features

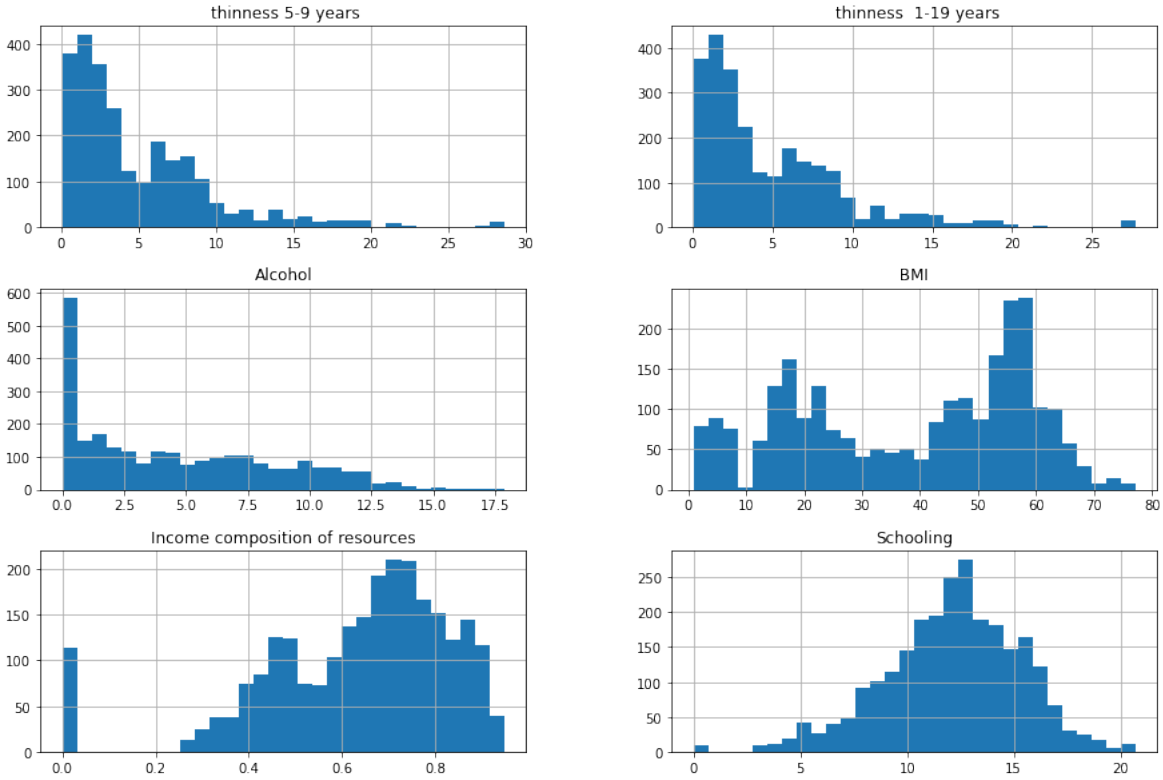


Table 2: Results of different regression models

Model	R squared	Num. of nonzero coefficients	Polynomial degree	The best alpha
Simple linear	0.8262	17	1	-
Polynomial	0.9084	153	2	-
Ridge	0.9219	488	3	10
Lasso	0.9244	137	3	0.005

the highest predictive performance together with relatively small number of parameters, which is better due to the small number of observations (2521).

Suggestions for the further analysis

I believe, that not normal distribution of the dependent variable could lower the prediction precision of the model. Hence I suggest finding a suitable transformation for Life expectancy in the further analysis. Moreover, it would be useful to collect additional information about Hepatitis B and GDP in developing countries like Vanuatu, Tonga, Togo, Cabo Werde, because such information is currently missing, and after that add Hepatiti B and GDP to the model, which could also increase the predictive power.