# Honors Peer-graded Assignment

## Supervised Machine Learning: Regression

**Michał Bożyk**

## 1. Dataset summary

Chosen was Life Expectancy (WHO) dataset that includes various factors that influence life expectancy (target feature). It consists of 2938 and has 22 features. Features describe for example whether person suffers from various diseases or consumes alcohol. They are mostly numerical however “Country” and “Status” features have types “object”. In some columns a vast number of null values was spotted. The main goal of this is to determine Life Expectancy values based on other features as well as to check which transformation, scaling and regularization techniques best suit the analyzed dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          2938 non-null   object 
 1   Year                             2938 non-null   int64  
 2   Status                           2938 non-null   object 
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64  
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64  
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64  
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years             2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

## 2. Plan for the data analysis

First, data will be cleansed, namely, NaN values will be replaced using forward fill. Next, “object” type features will be changed to numerical. Later, normally distribution of target variable will be examined by plotting the histogram and performing statistical test which will output p-value. Transformations will be performed if the normality won’t be sufficient:

- log,
- square root,
- BoxCox.

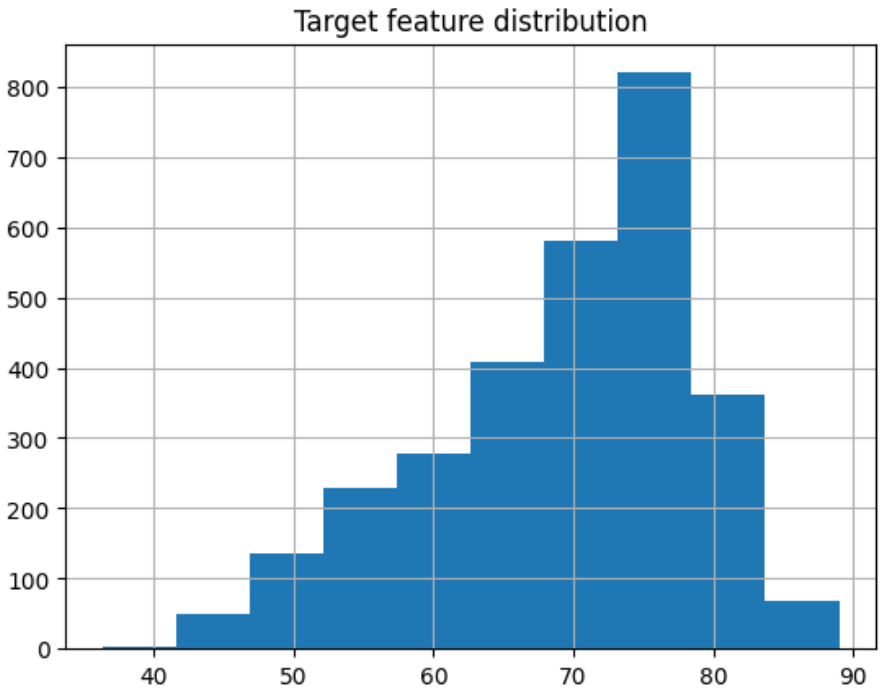
Next, data will be scaled. Three separate scaling techniques will be used and compared.:

- Standard Scaler,
- MinMaxScaler,
- MaxAbsScaler.

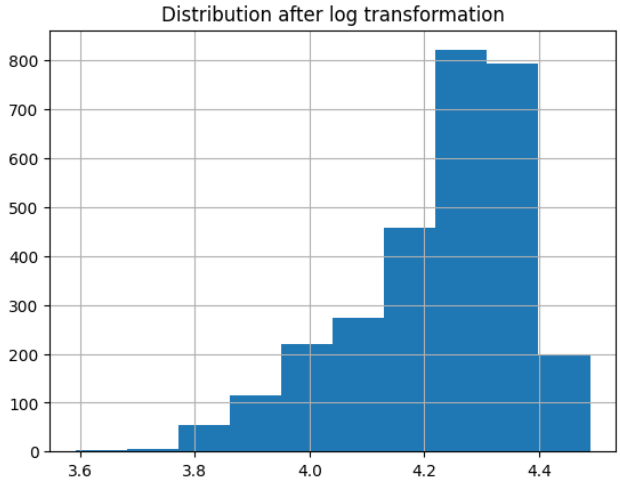
Two types of regularization will be applied: Lasso and Ridge. Different values of hyperparameters will be tested. Finally, Lasso and Ridge regression with optimized hyperparameters performance will be compared to each other.

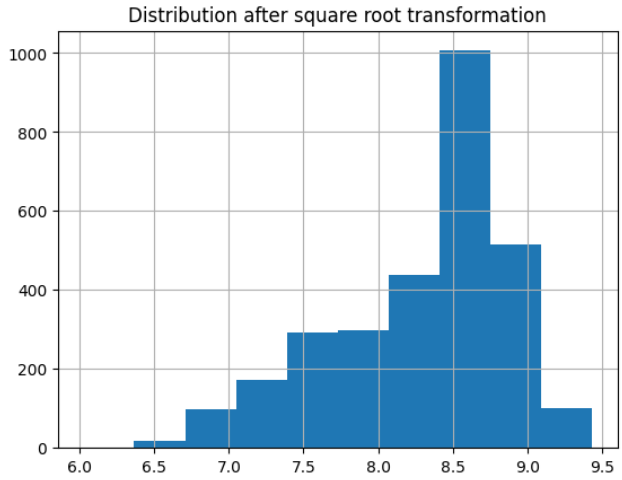
## 3. Comparison of transformation techniques

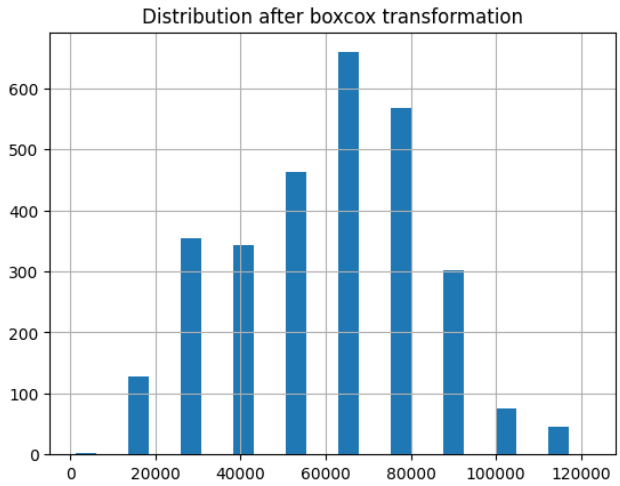
First, plotted was the distribution of target variable.



Applying log, sqrt and boxcox transformations yielded following results:



Results of D'Agostino K-squared test for normality for the non-transformed dataset:

**statistic=177.33620410175092,**

**pvalue=3.1040775189334384e-39**

The value of statistic suggests that that there is strong deviation from normality and the extremely small value of p-value indicates strong evidence against the null hypothesis that the data is normally distributed. Transformations will be applied and compared.

Next, the effectiveness of these transformation techniques was compared utilizing D'Agostino K-squared test for normality:

| Transformation | Statistics (deviation from normality) | p-value (significance of the null hypothesis) |
|---|---|---|
| log | 331.50266735668373 | 1.0354053755454299e-72 |
| sqrt | 240.082434121537 | 7.358034665336535e-53 |
| boxcox | 80.0253258744376 | 4.194896790656272e-18 |

Deviation from normality is still high even after applying transformation (in case of log and sqrt is even higher than it used to be). As it comes to p-value in each case is still incredibly low. It leads to the conclusion that relatively simple transformation methods are not efficient in terms of improving normality. It can be due to the high randomness of data what is characteristic for medical and health-related datasets. Further analysis was based on the original, non-transformed dataset.

# 4. Comparison of scaling techniques and degrees of polynomial features

Compared were StandardScaler, MinMaxScaler and MaxAbsScaler. All of the techniques yielded similar results as it comes to $R^2$ score and Mean Squared Error.

| Scaler | $R^2$ score | Mean Squared Error |
|---|---|---|
| StandardScaler | 0.8018505304833281 | 17.886983956038968 |
| MinMaxScaler | 0.8018505304833284 | 17.886983956038936 |
| MaxAbsScaler | 0.8018505304833285 | 17.88698395603893 |

Values for $R^2$ score and MSE are virtually the same meaning that scaling does not significantly influence the relationship between input features and target variables. However, by analyzing $R^2$ score it can be observed that 80% of variance is explained suggesting that linear regression was the right choice for this task. StandardScaler was chosen for further analysis.

Next, three  degrees of polynomial features were tested. The results are presented in the table.

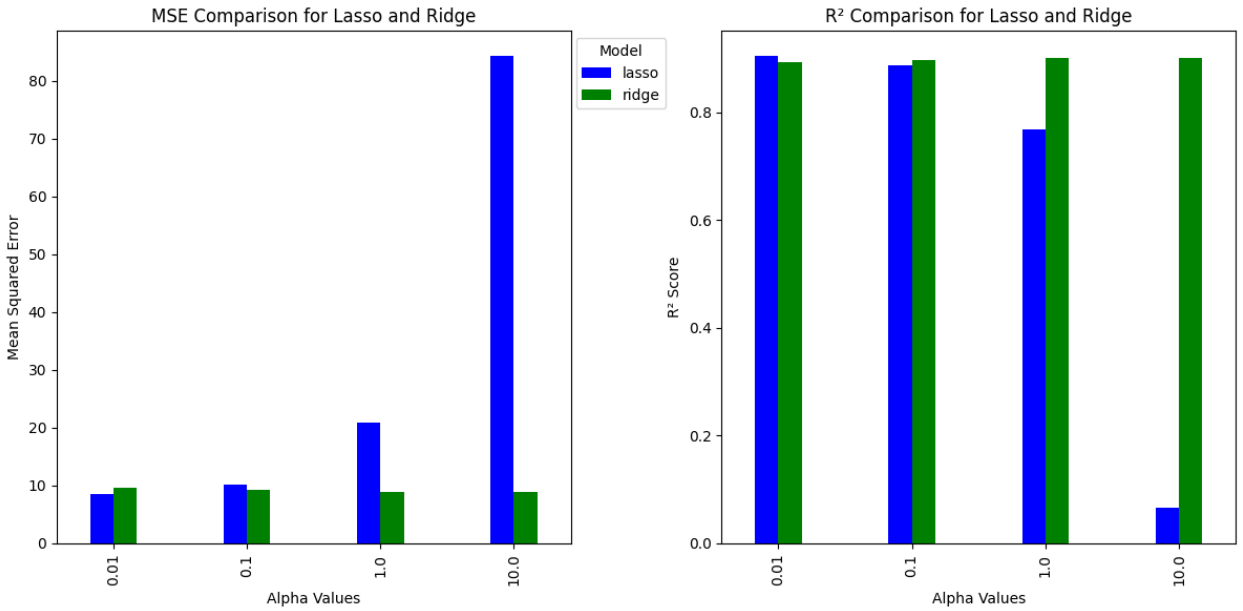
| Degree | $R^2$ score | Mean Squared Error |
|---|---|---|
| 1 | 0.8018505304833281 | 17.886983956038968 |
| 2 | 0.89025803810732 | 9.90642425875104 |
| 3 | 154532.77834922334 | 13949788.615336515 |

Degree 1 did not change anything what was expected since it is a simple linear model. As it comes to degree 2 the fit to the data is improved since the model captures more complex relationships between input features and the target. Degree 3 suggests overfitting due to abnormality of MSE and $R^2$ score. For further analysis degree 2 polynomial was chosen.

## 5. Comparison of regularization techniques and various alpha values

Lastly, two regularization techniques were tested and compared, namely Lasso and Ridge, together with various alpha values. The results were presented in form of the table as well as bar plot.

| Regularization | Alpha | $R^2$ score | Mean Squared Error |
|---|---|---|---|
| Lasso | 0.01 | 0.9059227209407579 | 8.492370861577712 |
| | 0.1 | 0.8872622769204516 | 10.176852094951373 |
| | 1 | 0.7691087712632807 | 20.842587739839175 |
| | 10 | 0.06530616254485921 | 84.37496055452172 |
| Ridge | 0.01 | 0.8933505663159597 | 9.62726124818904 |
| | 0.1 | 0.897686704390565 | 9.235837378315347 |
| | 1 | 0.9009357541825015 | 8.94254513967097 |
| | 10 | 0.9016175860693514 | 8.880996067394632 |



It turns out that lower values of alpha results in good performance for both Lasso and Ridge and the use of regularization techniques provides further improvements for linear regression. When it comes to larger values of alpha in case of Ridge regression the model still maintains reasonable performance, however Lasso regression's one is significantly worse than it used to be.

## 6. Conclusions

Analyzed was Life Expectancy WHO dataset. First, it was cleaned and NaN values were filled. Next, tested were various techniques to improve the robustness of linear regression model. It turned out that the use of transformation techniques to improve the normality of distribution is irrelevant, probably due to the high randomness of target feature. Various scaling techniques yielded similar results. However, when it comes to the degree of polynomial features the most suitable one was deemed to be 2. As for the regularization technique the most fitting seems to be Lasso alpha equal to 0.01. Model, after improvements, explain more than 90% of variance with relatively low MSE (close to 8.5). When it comes to future perspective of this project, the robustness of chosen parameters can be confirmed by Cross-validation.