Dedole Tommy
https://github.com/D-Tommy/IBM-Course

Pair-reviewed project : Predicting life-expectancy using linear and Lasso regression on WHO dataset.

Dataset Used :

- Life Expectancy (WHO) :
    - Statistical Analysis on factors influencing Life Expectancy

API : kaggle datasets download -d kumarajarshi/life-expectancy-who

direct link : Life Expectancy (WHO)

Context

Although there have been a lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on a data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations in 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 suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Content

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables were then divided into several broad categories:Immunization related factors, Mortality factors, Economical factors and Social factors.

Objective :

Use regression models in order to predict life expectancy.

- Interpretation is more important than prediction in this context given we want to know which factors are the most impacting the life expectancy of a population

Code used for this project is available on my github

1) Data cleaning and exploration

The dataset used here has 2938 entries for 22 features, containing 193 countries' information over the course of 15 years between 2000 and 2015 as mentioned above. (all variables can be explored through the provided code that includes a pandas-profiling report)

1.1) Feature names tuning.

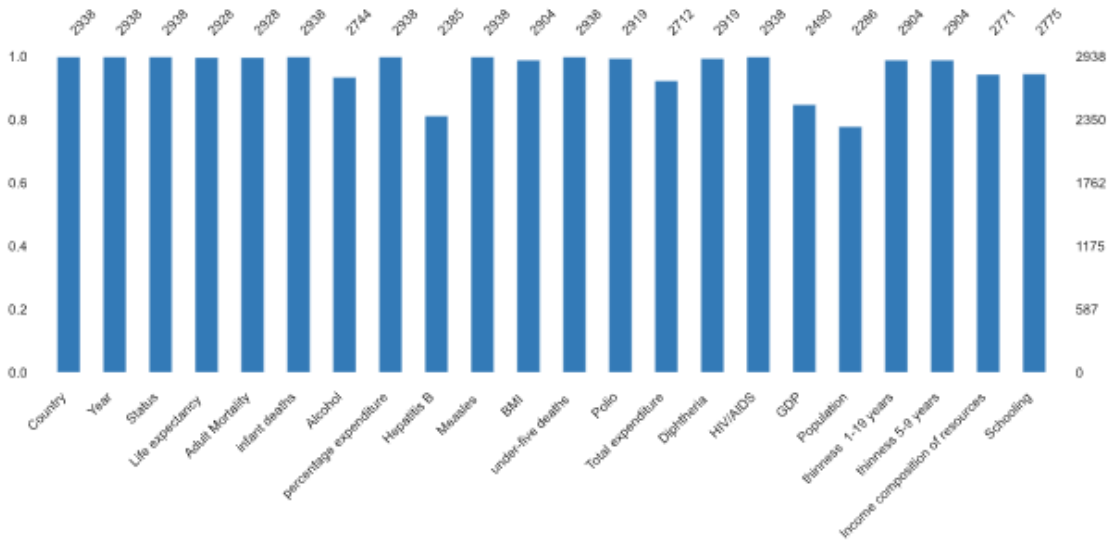
	0
0	country
1	year
2	status
3	life_expectancy
4	adult_mortality

The first issue about this dataset is column names : they contain upper cases / spaces (even some at the end and beginning) which makes coding tedious. Therefore, all column names have been passed through a function to make them consistent :

- Remove spaces at the beginning/end of it
- Replace inside space with '_'
- Remove upper cases.

1.2) Missing values and outliers.

There are a lot of missing values in this dataset :



A simple visualization of nullity by column.

It has been decided to remove the ‘population’ features since too many values were missing and the dataset already offers a lot of features. It would have been better to scrap wikipedia for missing data and implement them into this dataset but it would have also been way too much time consuming for little benefits.

1.2.1) Outliers

There are several outliers due to typing errors while filling the dataset :

percentage_expenditure	gdp
71.279624	584.259210
73.523582	612.696514
73.219243	631.744976
78.184215	669.959000
7.097109	63.537231
79.679367	553.328940

Example of typing mistake resulting in outliers.

These mistakes results in a lot of flawed rows :

	adult_mortality	infant_deaths	alcohol	percentage_expenditure	hepatitis_b
dtypes	float64	int64	float64	float64	float64
null_values	10	0	194	0	553
outliers	330	25	196	609	178

These outliers are detected, by country, the following way :

$$x < (Q25\% - 1.5 * IQR) \; or \; x > (Q75\% + 1.5 * IQR)$$

with IQR = Q75 - Q25

The outliers are then replaced by the mean value of this feature over the 15 years for the current country without them :

percentage_expenditure	gdp
71.279624	584.25921
73.523582	612.696514
73.219243	631.744976
78.184215	669.959
39.348286	382.24396
79.679367	553.32894

Outliers replaced by the mean of the country.

It would be more precise to do the mean of the 2 values around the outlier, however it would require more coding to take in count outliers on the first or last row of a country dataset.

1.2.2) Missing/null values

	adult_mortality	infant_deaths	alcohol	percentage_expenditure	hepatitis_b
dtypes	float64	int64	float64	float64	float64
null_values	10	0	194	0	553
outliers	330	25	196	609	178

This dataset contains many NaN values as well. There are two types of NaN in this case :

- 1) The whole feature is NaN for a country : the 15 rows (years) are empty.
- 2) Only a few of the rows are presenting a NaN for this feature.

Both cases have been handled differently :

- 1) Replace NaN with 0.
- 2) Replace NaN by the country's mean for this feature.

1.3) Encoding

The only feature that needs to be encoded is ‘status’ that only takes 2 categorical values.

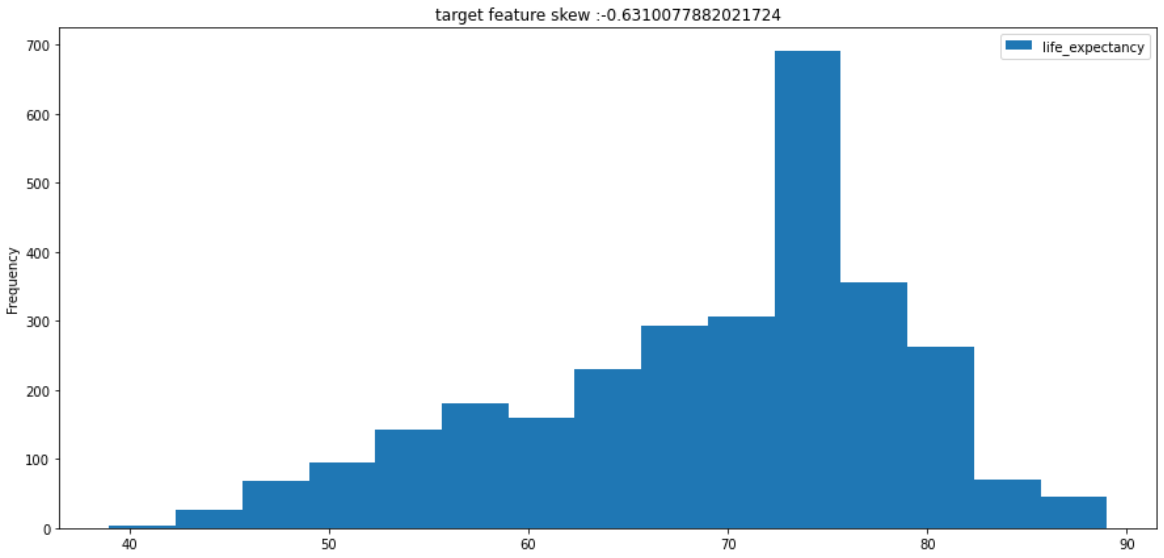
```
Entrée [8]:  1 df['status'].value_counts()

Out[8]: Developing    2416
        Developed      512
        Name: status, dtype: int64
```

```
Entrée [9]:  1 df['status'] = df['status'].map({'Developing':0, 'Developed':1})
             2 df['status'].value_counts()

Out[9]: 0    2416
        1     512
        Name: status, dtype: int64
```

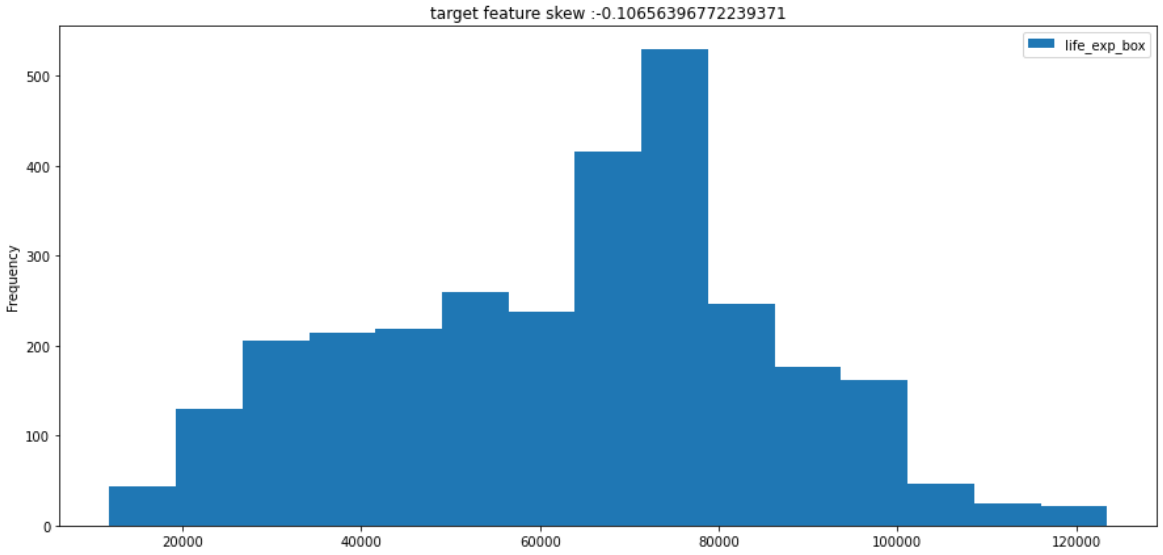
the ‘country’ feature will not be used in regression and therefore doesn’t need any encoding.

1.4) Target feature

The target feature here is the ‘life_expectancy’. Since it is planned to do regression, it is worth checking if it is correctly normally distributed :

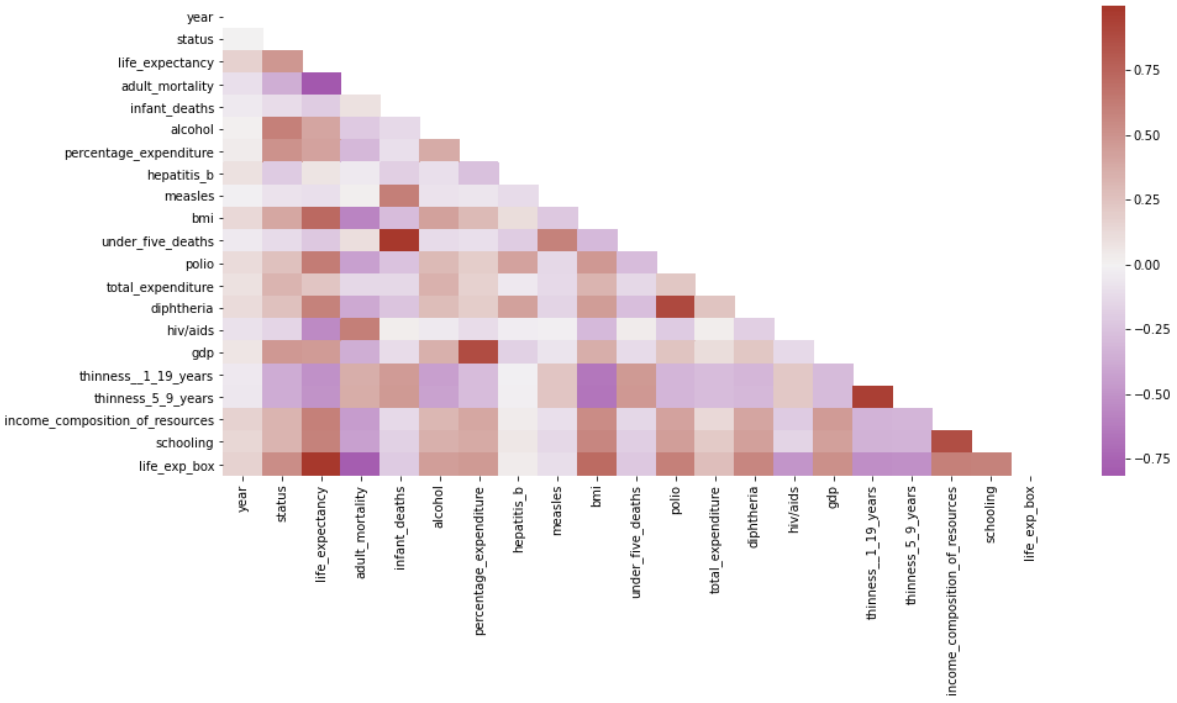


There’s a slight skew to the left. It can be enhanced using a box-cox transformation :



The skew of the target feature has been reduced significantly and make the target feature almost normally distributed, allowing ones to take it as target for regressions.

1.5) Features correlation (Pearson r)

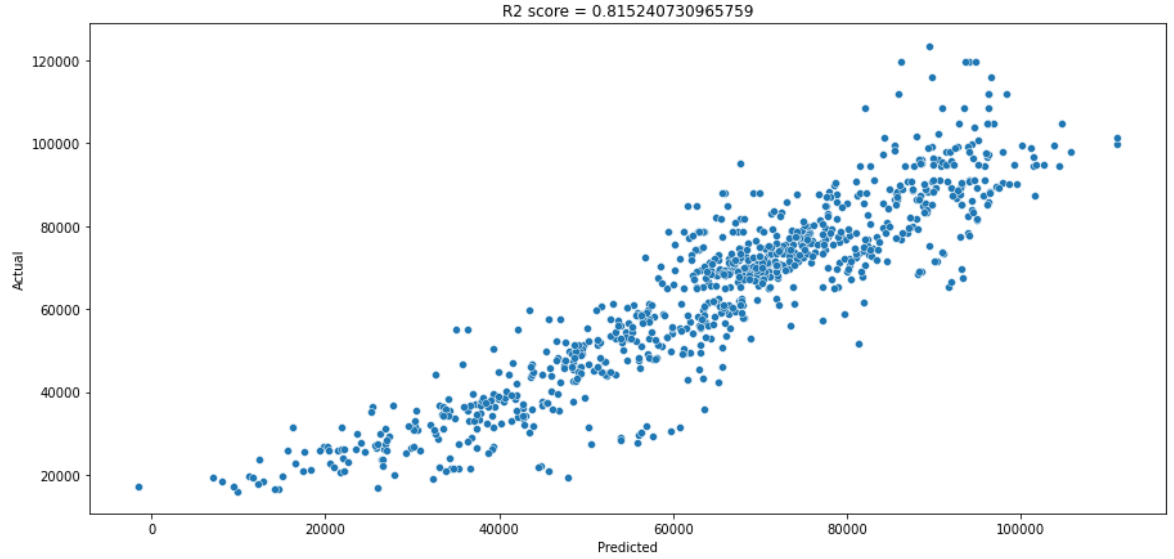


2) Regressions

The following features have been removed : country and year. They are of no interest when trying to compute the life expectancy and might even result in noisy features.

2.1) Basic linear regression

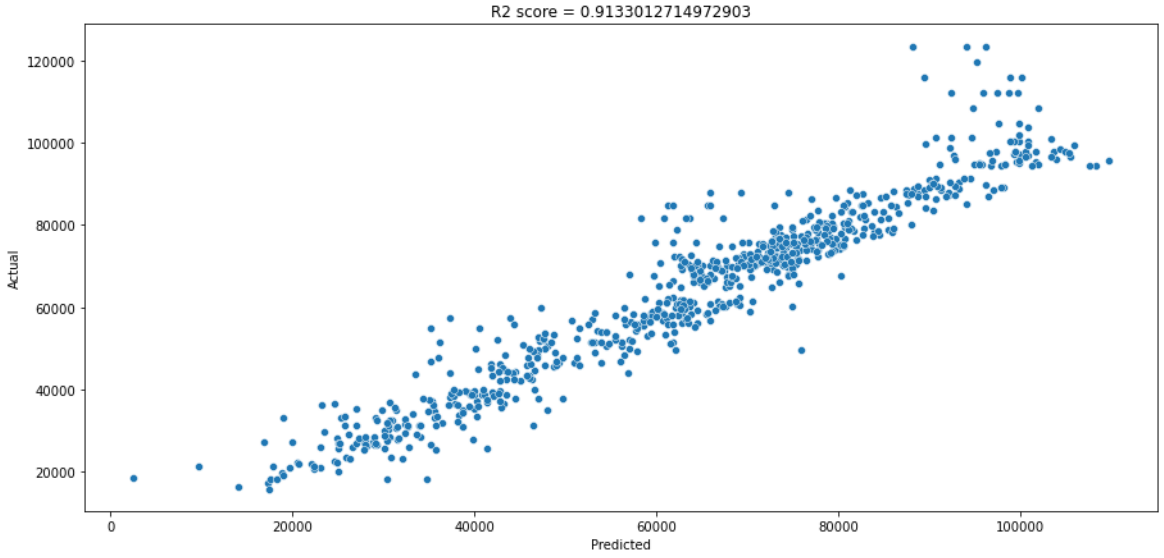
The first trained model is a basic linear regression on the cleaned dataset without any additional features. a train / test split with a factor of 0.3 is used.



Plot of predicted VS actual values obtained with a basic model, with a R2 score of 0.815. This is a decent model but it can still be improved.

2.2) Linear regression with polynomial features

Now polynomial features are added to the dataset, increasing the number of features to 189. The same train/test split is used.



Adding the polynomial features significantly reduced the variance of our model and its bias making it far better than the previous one.

However, the objective here is to have a model that provides the user with high interpretability and it is not the case for a linear regression using almost 200 features.

2.3) Lasso regression with polynomial features

Once more, polynomial features are used and then scale our data using a standard scaler. The best regularization factor α is chosen using as grid search :

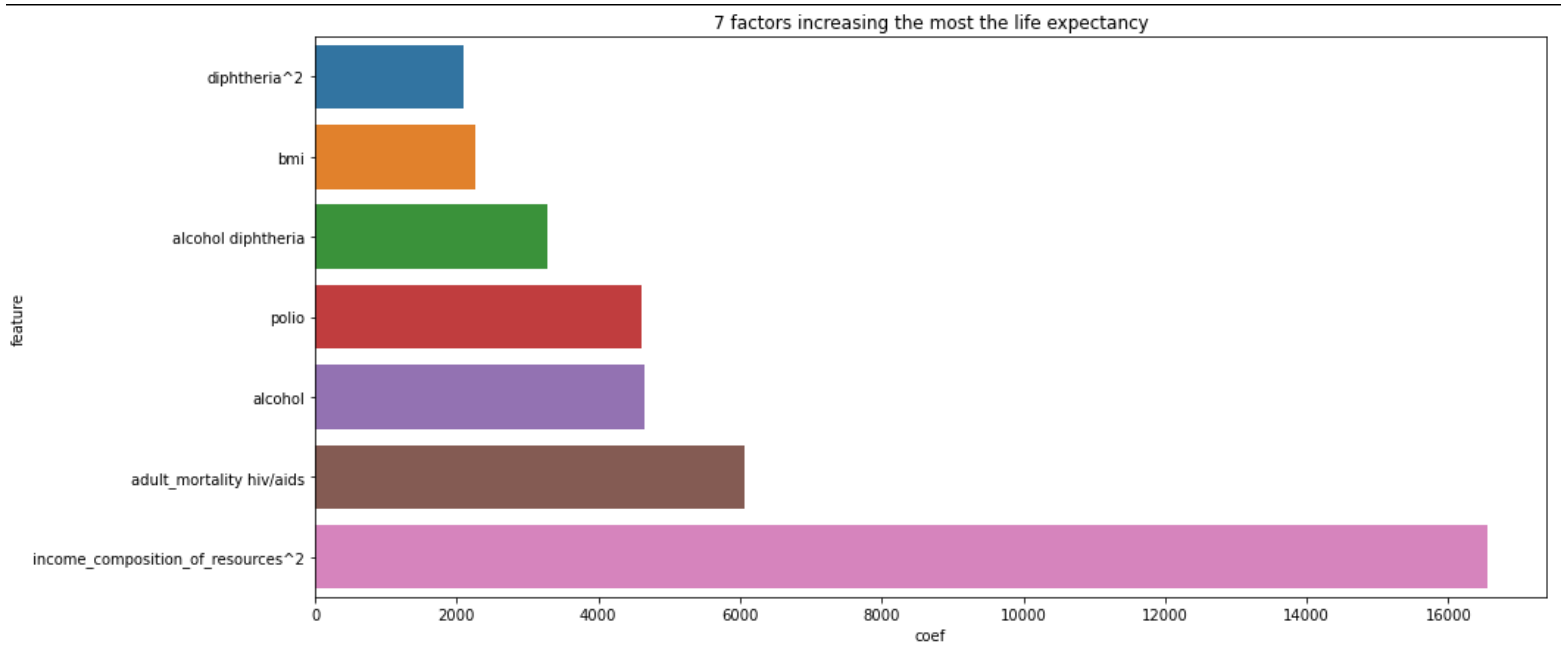
Entrée [36]:

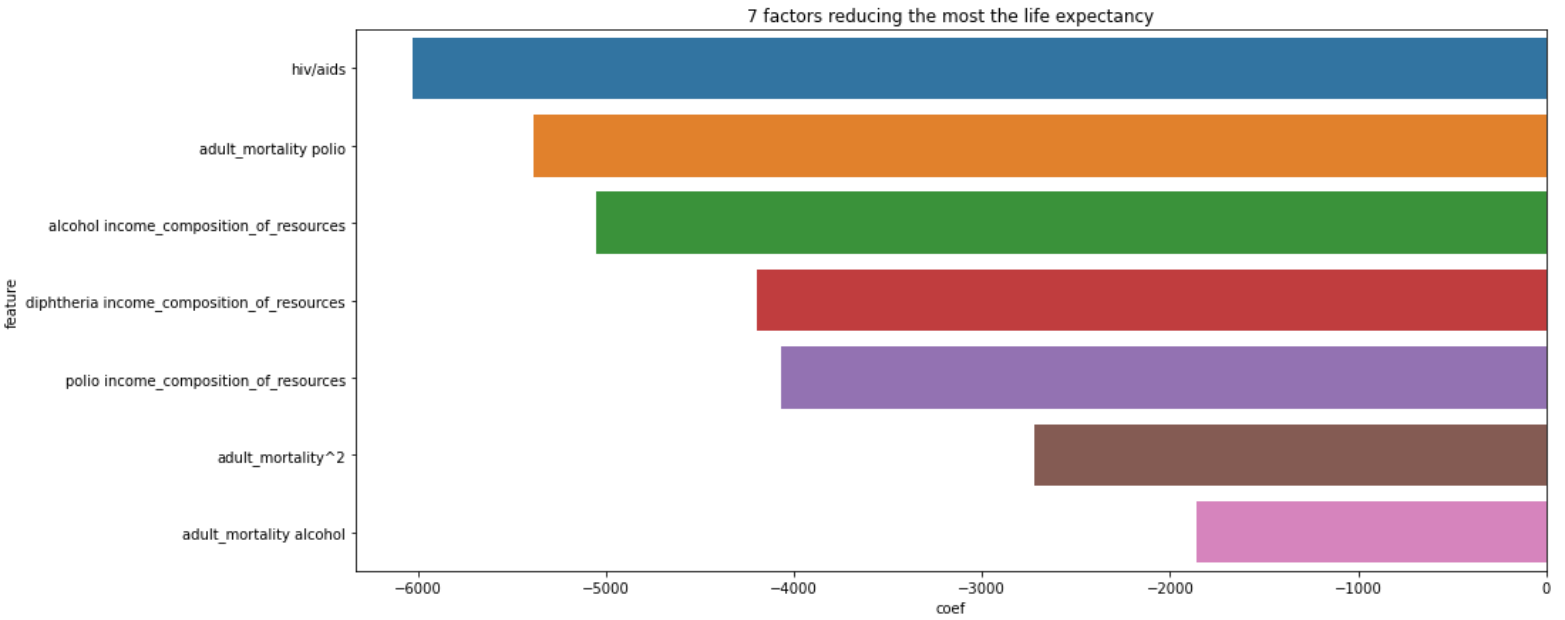
```
alphas = [0.01, 0.1, 1, 10, 100, 150]
lasso = Lasso(max_iter = 10000)

grid = GridSearchCV(estimator = lasso,
                    param_grid = {'alpha':alphas},
                    scoring = 'r2',
                    cv = 5,
                    )
grid.fit(Xscaled, y)
alpha = grid.best_params_
score = grid.best_score_

print(f'best r2 score obtained : {score} for alpha = {alpha}')
```

```
best r2 score obtained : 0.8683065591851042 for alpha = {'alpha': 100}
```

α = 100 is a high value for the regularization factor that ensures the user with a lot of 0 coefficients.
Moreover this model performs better than the initial linear regression. But slightly worse than the one using all of the features.

However, since this model is built to have a high interpretability, it is possible to verify it by checking how fast the coefficients are decreasing, and the most important ones:





Unsurprisingly, the factor that has the highest positive impact on life expectancy is the income of a population, and the one that reduces it the most is the amount of HIV/AIDS.

However, a lot of the features used here are somehow correlated together (e.g : alcohol consumption and adult mortality) and these results should not be considered as serious enough to base any decision on it because of this. These findings remains the outcome of a quick project over only 2 days.

Conclusion

The most interesting model for our study here is the Lasso regression based on polynomial features which is the perfect tradeoff between interpretability and performance.
Even though the dataset contained a lot of flawed rows, data cleaned allowed us to use almost the entirety of it.
This dataset could also be used in unsupervised learning by trying to cluster populations by their health features and verify if these clusters present significantly different life expectancy.