Life Expectancy Prediction

Exploratory data analysis: Dataset

Dataset:
- Public dataset made available by the World Health Organization (WHO)
- 21 predictor variables and 2938 observations
- The dataset includes features belonging to different macrocategories:

Immunization factors:
- Hepatitis B
- Polio
- Diphteria

Economic factors:
- GDP
- Total expenditure
- Income composition of resources
- Percentage expenditure

Mortality factors:
- Thinness 5-9 years
- Thinness 1-19 years
- Adult mortality
- HIV/AIDS
- Measles
- Infant deaths
- Under-five deaths

Social factors:
- Population
- Alcohol
- Schooling
- BMI

Other factors:
- Country
- Year
- Status

Output variable:
- Life expectancy

Exploratory data analysis: Outliers handling

Problem:

- The dataset is heavily affected by outliers
- It is unrealistic that the Population for a specific country drops by a factor of 10/100/1000 from one year to the following one
- Similar considerations apply to other features, such as Polio, Measles, GDP…

Premise:

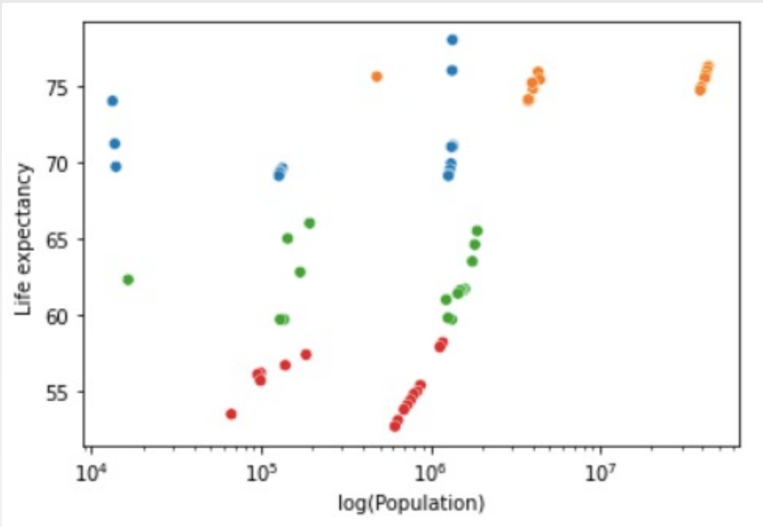
- Features are not expected to change their order of magnitude during the years if they are referred to the same country

How was it solved:

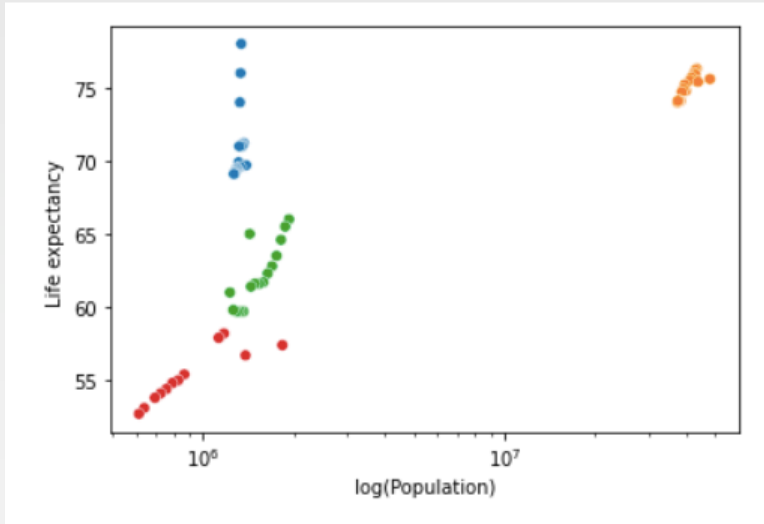
- Identifying the most common order of magnitude for each feature and each country
- Multiplying/Dividing by a factor of 10*k (with -3<k<3), so that all the features for the same country have the same order of magnitude

Exploratory data analysis: Outliers handling

Pre-outliers handling



Post-outliers handling



Exploratory data analysis: NaN handling

Problem:

- The dataset is affected by missing values (14 out of the 21 features)
- The percentage of NaN for different features spans from < 1% to 22 % of the total number of observations

Premise:

- Features are not expected to change a lot their values during the years if they are referred to the same country

How was it solved:

- For each feature, if there is at least one value which is not NaN for a country, the mean of these values is to replace NaN for that specific country
- For each feature, if all the values are NaN, iterative imputation is used to address NaN

Exploratory data analysis: Correlation between variables

Correlation among predictor variables:

- The dataset is affected by multicollinearity. Pearson correlation coefficients were all above 0.8 for the following features:
  - Population - under five deaths - infant deaths -> Population and infant deaths are dropped
  - Thinness 1-19 years - Thinness 5-19 years -> Thinness 5-19 years is dropped
  - Diphteria - Polio -> Diphteria is dropped

Correlation between predictor and output variables:

Before outliers handling - top correlated variables

	R^2
Schooling	0.77
Income composition of resources	0.74
BMI	0.57
HIV/AIDS	-0.56
Adult Mortality	-0.70

After outliers handling - top correlated variables

	R^2
Income composition of resources	0.86
BMI	0.71
Polio	0.62
Schooling	0.61
Adult Mortality	-0.90

Exploratory data analysis: Categorical variables and Dataset splitting

Transformation of categorical variables:

- 2 categorical variables (Status and Country), respectively having 2 and 193 unique values
- One hot encoding was performed for both categorical variables; this ended up increasing the number of features from 18 to 197

Splitting and scaling the dataset:

- 80% training set, and 20% in the test set
- K-fold validation performed on the dataset (10 folds)
- Data scaled using the standard scaler

Models: Linear regression

Metrics used:

- Mean squared error (MSE)
- Coefficient of determination (R^2)

Models used:

- Vanilla linear regression
- Ridge linear regression
- Lasso linear regression

Summary of the results obtained:

	Vanilla linear regression	Ridge regression	Lasso regression
R^2 – training set	0.966	0.965	0.965
R^2 – test set	0.954	0.957	0.956
R^2 – k-fold	0.953	0.937	0.954
MSE – training set	3.032	3.121	3.068
MSE – test set	4.314	4.053	4.065
MSE – k-fold	4.287	5.739	4.135

Models: Linear regression

Hyperparameter tuning:

- Ridge regression: alpha=10 np.geomspace(0.0001, 100, 7)
- Lasso regression: alpha=0.001 np.geomspace(0.0001, 100, 7)

Comments on the results:

- Similar performances for the three models in terms of MSE and R^2 on the test set
- Training and test MSE are similar, while there seems to be some bias (training error can be improved)
    - The presence of bias is consistent with the fact that simple models are used (regression)
    - More complex models need to be implemented

⬇

Ensemble models

Models: Ensemble models

Models used:

- Bagging
- Random forest
- XGBoost

Hyperparameter tuning:

- Bagging:
  - Number of estimators: 120 — n in range(50,131,10)
  - Max depth base estimator: 80 — n in range(30,81,10)

- Random forest:
  - Number of estimators: 130 — n in range(50,131,10)
  - Max depth base estimator: 30 — n in range(1,41,10)
  - Max features: auto — auto, sqrt

- XGBoost:
  - Number of estimators: 100 — n in range(40,101,10)
  - Max depth base estimator: 31 — n in range(1,41,10)
  - Learning rate: 0.1 — n in np.geomspace(0.001,1,4)

Models: Ensemble models

Models used:

- Bagging
- Random forest
- XGBoost

Summary of the results obtained:

	Bagging	Random forest	XGBoost
R^2 – training set	0.994	0.996	1.000
R^2 – test set	0.949	0.969	0.969
R^2 – k-fold	0.963	0.969	0.970
MSE – training set	0.507	0.349	0.001
MSE – test set	4.135	2.817	2.816
MSE – k-fold	3.349	2.815	2.689

Models: Ensemble models

Models used:

- Random forest performing EDA
- Random forest without performing EDA

Summary of the results obtained:

	Random forest with EDA	Random forest without EDA
R^2 – training set	0.996	0.995
R^2 – test set	0.969	0.963
R^2 – k-fold	0.969	0.962
MSE – training set	0.349	0.446
MSE – test set	2.817	3.287
MSE – k-fold	2.815	3.452

Conclusions

- EDA has significally improved the performance of the model (Random Forest MSE on the test set dimished by 12%)

- Ensemble models showed much better performance with respect to linear regression. The best ensemble model (XGBoost) allowed to reduce MSE on the test set up to 31% with respect to the best regression model (Ridge).