Dying to know!

Predicting life expectancy using WHOs data

By Eduardo Beltrán Herrera

IBM Machine Learning Specialization

01/10/2020

Introduction

To live a long, fulfilling life is perhaps the biggest desire of all human beings. While measuring the quality of one's existence is impossible, the duration of said existence is very much quantifiable and measurable, and we can try to predict it with some accuracy.

The following report utilizes life expectancy data of various countries across many years to try and train a linear regression model to estimate the life expectancy of said population. The dataset was extracted from Kaggle.com (https://www.kaggle.com/kumarajarshi/life-expectancy-who/discussion).

The table in its raw form contains 2,938 observations of **life expectancy** (in years) on different countries across many years, with 21 other measures of each location:

1. **Country**
2. **Year**
3. **Status** (developed or developing)
4. **Adult mortality** (number of deaths before 60 years of age per 1000 population)
5. **Infant deaths** (number of deaths per 1000 population)
6. **Alcohol** (consumption in liters per capita)
7. **Percentage expenditure** (expenditure on health as a percentage of GDP)
8. **Hepatitis B** (percentage of immunization coverage)
9. **Measles** (reported cases per 1000 population)
10. **BMI** (average body mass index of the entire population)
11. **Under-five deaths** (per 1000 population)
12. **Polio** (percentage of immunization coverage)
13. **Total expenditure** (percentage of government expenditure on health from budget)
14. **Diphtheria** (percentage of immunization coverage)
15. **HIV/AIDS** (deaths per 1000 live births)
16. **GDP** (gross domestic product)
17. **Population**
18. **Thinness 1-19 years** (prevalence of thinness among children and adolescents)
19. **Thinness 5-9 years** (prevalence of thinness among children)
20. **Income composition** (Human Development Index)
21. **Schooling** (number of years in school)

The intention of the applied models will be purely **predictive**, and thus the interpretation of the models will not be considered in the final analysis. The accuracy of the different models will be compared with the R^2 metric.

Data cleansing

A quick glance at the data shows us there a significant number of missing values in certain columns, particularly in the Population column:

#	Column	Non-Null Count	Dtype
0	Year	2938 non-null	int64
1	Life expectancy	2928 non-null	float64
2	Adult Mortality	2928 non-null	float64
3	infant deaths	2938 non-null	int64
4	Alcohol	2744 non-null	float64
5	percentage expenditure	2938 non-null	float64
6	Hepatitis B	2385 non-null	float64
7	Measles	2938 non-null	int64
8	BMI	2904 non-null	float64
9	under-five deaths	2938 non-null	int64
10	Polio	2919 non-null	float64
11	Total expenditure	2712 non-null	float64
12	Diphtheria	2919 non-null	float64
13	HIV/AIDS	2938 non-null	float64
14	GDP	2490 non-null	float64
15	Population	2286 non-null	float64
16	thinness 1-19 years	2904 non-null	float64
17	thinness 5-9 years	2904 non-null	float64
18	Income composition of resources	2771 non-null	float64
19	Schooling	2775 non-null	float64
20	Status_Developing	2938 non-null	uint8

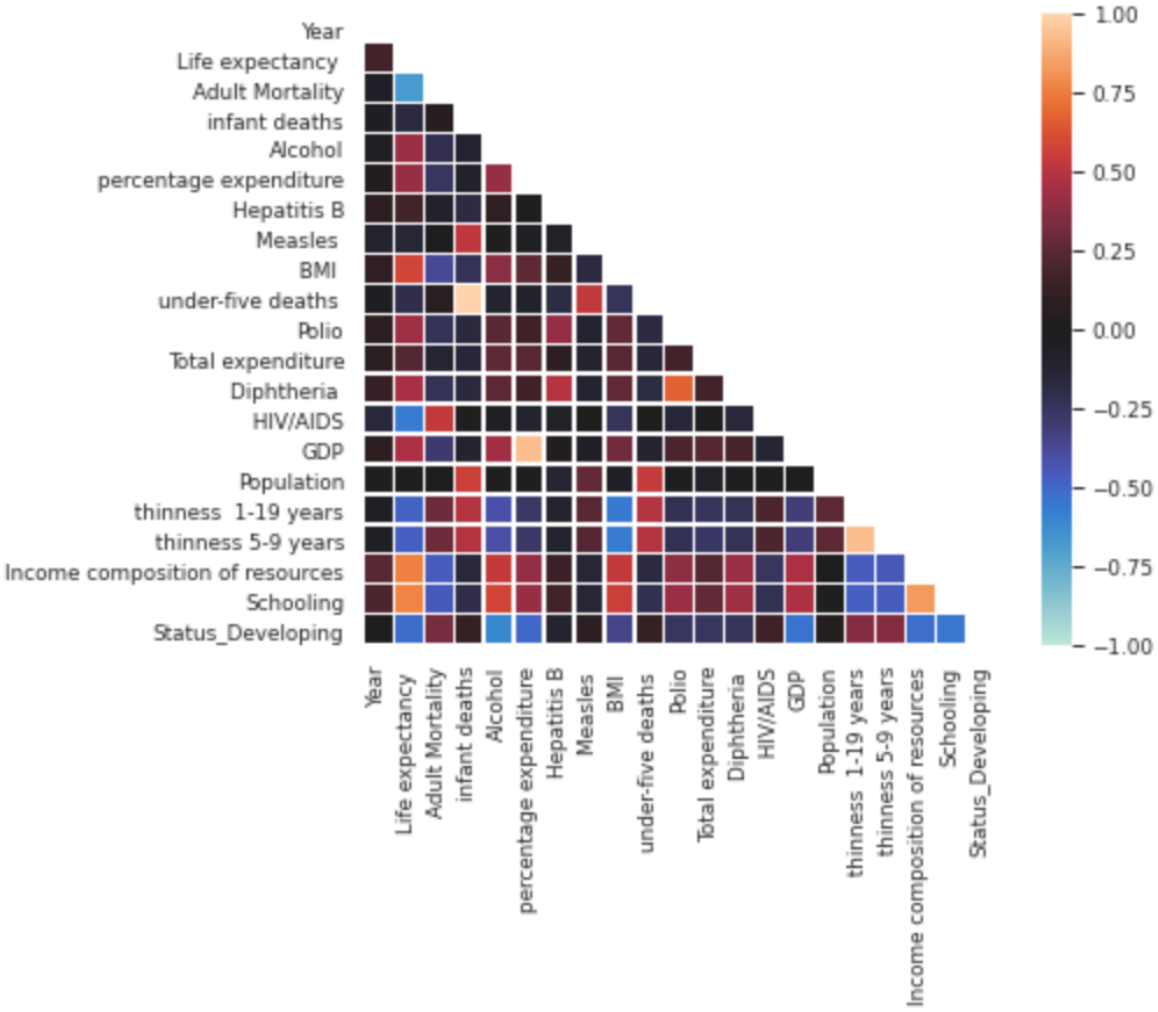
All features with count of non-missing values and data type

Further analysis shows us that certain countries lack that feature, and so all missing registries were eliminated from the dataset. All other missing values were replaced with the average value of their respective column, so that we can preserve as many rows as possible.

A heatmap with the correlation coefficient was created to identify redundant or highly correlated variables. The highest correlated variables observed were:

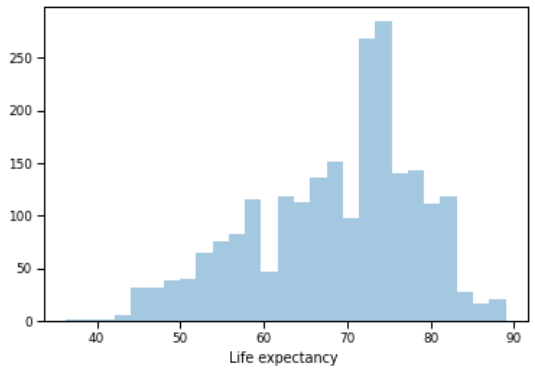
- **‘under-five deaths’ – ‘infant deaths’**
- **‘GDP’ – ‘percentage expenditure’**
- **‘thinness 5-9 years’ – ‘thinness 1-19 years’**

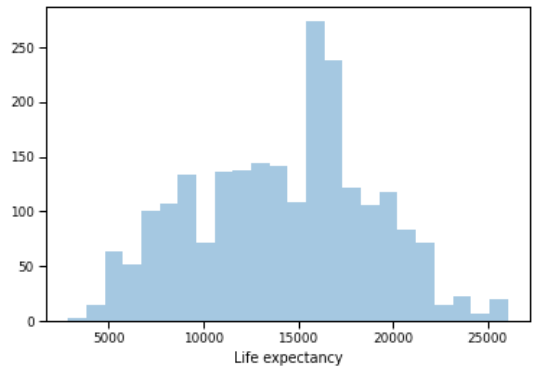
The second variable in each pair was eliminated from the set so the model can be simplified.



Heatmap of correlation coefficient between features

Finally, the target variable was tested for normality and transformed accordingly. Unfortunately, no transformation could pass the normality test. Tre transformation was then discarded as to not complicate the modeling process.





Target variable before and after 'boxcox' transformation. Neither p-value was high enough to justify transformation

Regression Modeling

The following models were trained and tested against the dataset. All models were evaluated with cross-validation with 10 folds. Hyperparameters for Ridge and LASSO models were obtained via the *GridSearch* algorithm:

Model	Description	Polynomial Degree	alpha	R^2 Score
LR	Standard Linear Regression	1	-	0.837
LR with PF	Standard Linear Regression model with polynomial features	2	-	0.916
Ridge	Ridge Regression with optimized hyperparameters	2	19.22	0.92
LASSO	Lasso regression with optimized hyperparameters	2	0.014	0.92

Summary of models trained and tested on set

Analysis

The biggest contributor to a more accurate model was the addition of the polynomial degrees, improving the R^2 score from 0.837 to 0.916. Applying regularization techniques improved on the model a bit more.

The model might be further improved with more data from the coming years, as the current dataset might be a bit small. Overall, the model has a satisfactory response. An even more accurate model could be constructed with other environmental variables, such as air quality of the country or average temperature throughout the year.

Conclusion

Regression models are very easy to implement using programming skills; the real challenge lies in the cleaning of the data and the feature engineering: choosing the right features to keep and encoding every

variable in the correct way will make for better predictive models no matter the way they are implemented.

Even so, some regression models are better than others. Regularized implementations tend to always show an improvement over their non-regularized counterpart. The complexity of the model is a double-edged sword that needs to be carefully examined to avoid overfitting to our data. Doing train-test splits or cross validation is crucial for evaluating any model.