Life Expectancy and Fertility Rates Among Nations

REGRESSION ANALYSIS OF FACTORS CONTRIBUTING TO LONGER LIVES AND LESS CHILDREN

Paul Gagliardi | Data Science Tools 2 | 11/19/2021

Introduction

Life expectancy and fertility rates are metrics that can be universally agreed on as something that we want to improve. Both these metrics are used as benchmarks to evaluate the development of a country as well as the strength of their healthcare system. To improve life expectancy and fertility rates we will need to dive into the factors that impact them. **This project aims to determine which factors or features in the dataset contribute the most to life expectancy and fertility rates.**

The method to be used in this paper will be regression analysis. These models can help us understand which features or input variables have the largest impact on the output variables of Life Expectancy and Fertility Rates, and are often by researchers trying to gauge the causal relationships between an input and output variable (i.e. inversely proportional, proportional, no or little relationship). **As such we will also be examining the strength of several different regression models and attempting to see which model is the best fit for this analysis.**

The models to be used are:

- Linear Regression
- RidgeCV Regressoin
- Lasso Regression
- Random Forrest Regressor
- XGBoost Regressor

Data

The dataset used for this analysis was published and made available on the popular Data Science website Kaggle, and aggregated by Kacper Kalinowski. It is available here: https://www.kaggle.com/kacperk77/life-expectancy. The data originally was collected from https://data.worldbank.org/ and is reflective of metrics taken in 2014 across 138 countries.

This overall dataset consists of two overlapping datasets:

- Dataset 1: 63 countries, 20 variables, and 17 input variables.
- Dataset 2: 138 countries, 13 variables, and 10 input variables

VARIABLES

- **Fertilityrate (Fertility rate, total (births per woman) 2014)**
- **Lifeexp (Life expectancy at birth, total (years) 2014)**
- Country (String description of country)
- *Literacyrate (Literacy rate, adult total (% of people ages 15 and above) 2014)
- *Homicidiesper100k (Homicidies per 100k people 2014)
- *Electricity (Electric power consumption (kWh per capita) 2014)

- Schooling (Ave number of years of Schooling (years) 2014)
- HIV.AIDS (Deaths per 1 000 live births HIV/AIDS (0-4 years) 2014)
- Status (Economic development status of country)
- Wateraccess (Access to improved water sources (% of total population with access 2014)
- Tuberculosis (Incidence of tuberculosis per 100,000 people 2014)
- Inflation (Inflation, consumer prices (annual %) 2014)
- Healthexppercapita (Average health expenditure per capita, PPP 2005-2014)
- *Internet (Individuals using the Internet (% of population) 2014)
- *Gdppercapita (GDP per capita, PPP (current international $) 2014)
- CO2 (Average CO2 emissions (metric tons per capita) 2005-2014)
- *Forest (Forest area (% of land area) 2014)
- *Urbanpop (Urban population 2014)
- Urbanpopgrowth (Average urban population growth (annual %) 2005-2014)
- Leastdeveloped (1 = country is considered as least developed; 0 = country is considered as developing or developed)

Here the features that have a star are features that only exist in Dataset 1, the dataset with 20 columns, but not in Dataset 2, the set with more countries. Output variables are **bolded and underlined.**

PREPROCESSING

This data was not missing any values, but did require some partitioning and changes to be ready for regression. 'Leastdeveloped' and 'Status' were redundant categorical variables that contained no objective numerical information. Since the status of a country's development carries no real word input that could rathe be thought of as a categorical output, these two features were removed from the analysis. Additionally, 'Country' was removed as it was a string that carried no relevant information for the regression.

The remaining input variables were all continuous, however they had large variance in mean values, so I opted to standardize all values. This can be useful in determining the most impactful factors in a regression, as the largest coefficients will then indicate the most significant features (as long as p-values are met).

	Schooling	HIV.AIDS	wateraccess	tuberculosis	inflation	healthexppercapita	fertilityrate	lifeexp	CO2	urbanpopgrowth	leastdeveloped
count	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000
mean	13.044203	0.643478	88.554348	121.607681	3.816283	1111.596390	2.832897	71.615724	4.817753	2.387124	0.253623
std	2.939835	1.311708	14.732248	162.039059	5.065714	1347.764785	1.395983	8.202395	6.832746	1.980082	0.436669
min	5.300000	0.100000	40.000000	0.760000	-1.509245	34.289633	1.230000	50.621000	0.027033	-1.342915	0.000000
25%	10.825000	0.100000	81.725000	14.000000	0.860427	158.076743	1.742500	66.151250	0.587155	0.932280	0.000000
50%	13.100000	0.100000	95.350000	53.000000	2.704123	545.677923	2.302000	73.752500	2.396420	2.211662	0.000000
75%	15.200000	0.400000	99.675000	166.500000	5.208063	1372.105575	3.752000	77.151037	6.396539	3.687510	0.750000
max	20.400000	9.400000	100.000000	852.000000	36.906643	5812.713900	7.338000	83.587805	47.488879	11.588452	1.000000

Dataset 2 example

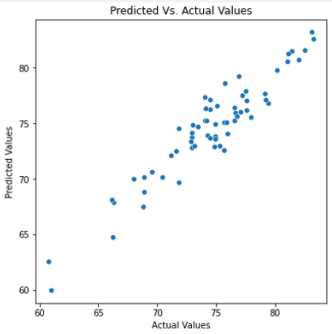
Dataset Selection and Regression Requirements

Before running all models on a dataset for both outputs, it was important to pick a dataset between the two available that would best set up future models for success. To do this, I ran a simple OLS Regression using statsmodel to determine if there was any significant information gain to be had by including the additional variables in Dataset 1. Here a "significant factor" is a factor that has a p-value of < 0.06.

LIFE EXPECTANCY

DATASET 1

- Significant factors:
  - HIV.AIDS
  - Tuberculosis
  - Healthexppercapita
  - CO2

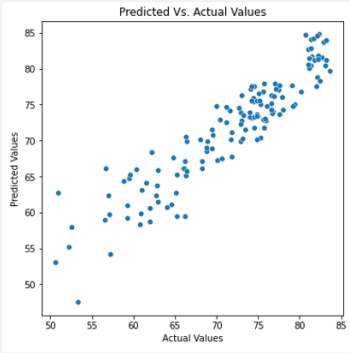


EVALUATION METRICS
R2: 0.89774
MAE: 1.25472
MSE: 2.27055
RMSE: 1.50683

AIC: 262.4
BIC: 296.7

DATASET 2

- Significant factors:
  - Schooling
  - HIV.AIDS
  - wateraccess
  - Healthexppercapita



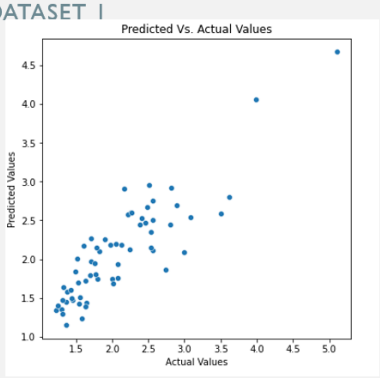
EVALUATION METRICS
R2: 0.86568
MAE: 2.32598
MSE: 8.97154
RMSE: 2.99525

AIC: 712.4
BIC: 738.8

LIFE EXPECTANCY

DATASET 1

- Significant factors:
  - HIV.AIDS
  - wateraccess
  - urbanpopgrowth

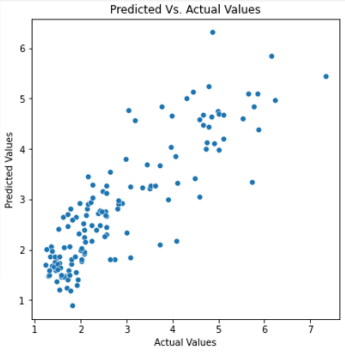


EVALUATION METRICS
R2: 0.76547
MAE: 0.27123
MSE: 0.12303
RMSE: 0.35076

AIC: 78.78
BIC: 113.1

DATASET 2

Significant factors:
- Schooling
- HIV.AIDS
- wateraccess
- tuberculosis
- CO2
- urbanpopgrowth

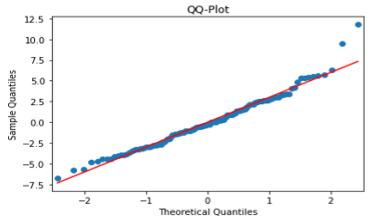


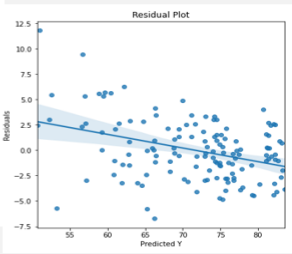
EVALUATION METRICS
R2: 0.75144
MAE: 0.52361
MSE: 0.48087
RMSE: 0.69345

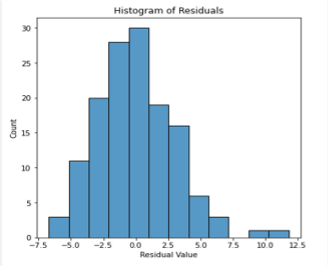
AIC: 308.6
BIC: 334.9

While the R-Squared score and the AIC/BIC is better for Dataset 1, Dataset 2 finds just as many or more "significant factors" than Dataset 1 which has more features. To avoid noise due to insignificant features, I chose to work with Dataset 2 as it has more countries which in my mind translated to more significant data overall.

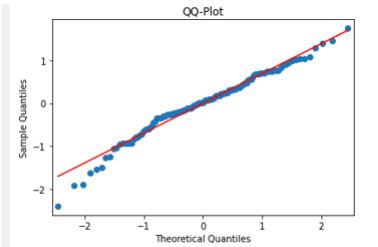
Life Expectancy Residuals

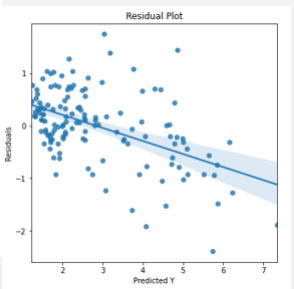


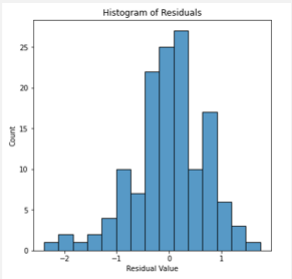




Fertility Rate Residuals







Regression Assumptions

The Residuals all obtained here are from an OLS regression on Dataset 2.

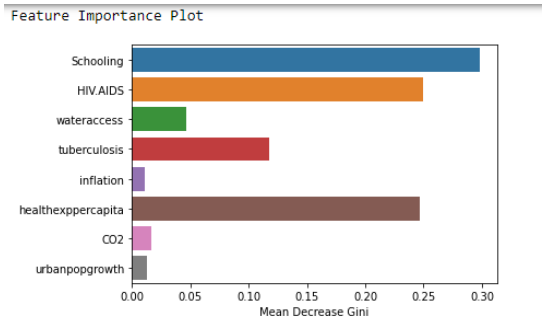
- Linear Relationship: Met
  - Tested via pairplot, see appendix
- Normality of Residuals: Met
  - Tested via QQ-Plots and Residual Histograms (left)
- Independence of residuals: Met
  - Assumed due to no time series data
- Homoscedasicity: Violated
  - The Fitted vs Residuals plot for both output variables show systematic decrease with higher output Y
  - Additionally, the dispersion is tightest at high Y for Life Expectancy and low Y for Fertility rates
  - Likely that some variables are not independent of eachother, leading to Heteroscedasticity. **Not deemed too big of an impact for this analysis**

Model Selection and Results

Life Expectancy

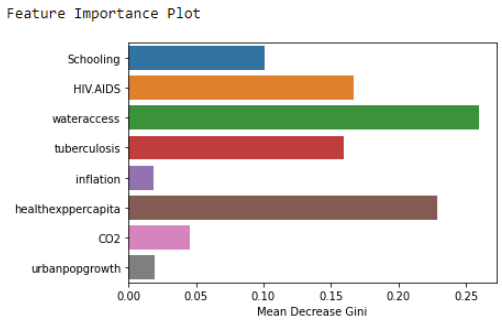
Dataset 2 was split into Train and Test sets for model selection, with 20% of data stored in test. Out of the 5 models ran, **Random Forest and XGBoost Regressors** had the highest R-Squared scores and lowest Mean Squared Errors. These models agreed that 'HIV.AIDS' and 'Healthexppercapita' were important features, but disagreed on some others.

Random Forest



EVALUATION METRICS
R2: 0.93271
MAE: 1.60593
MSE: 4.8943
RMSE: 2.21231

XGBoost



EVALUATION METRICS
R2: 0.92995
MAE: 1.76755
MSE: 5.09462
RMSE: 2.25713

Fertility Rate

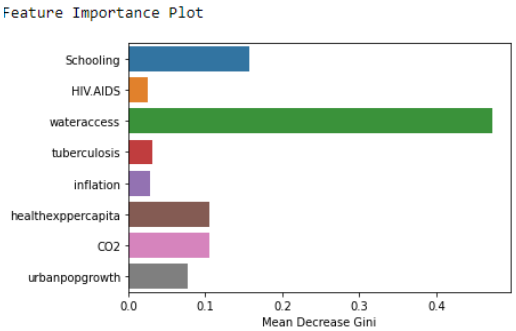
Out of the 5 models ran, **Random Forest and RidgeCV Regressors** had the highest R-Squared scores and lowest Mean Squared Errors.

Ridge CV

COEFFICIENTS
Intercept: 2.8085
Schooling: -0.346136
HIV.AIDS: 0.182732
wateraccess: -0.524717
tuberculosis: -0.089533
inflation: -0.01353
healthexppercapita: -0.01301
CO2: -0.263543
urbanpopgrowth: 0.351909

EVALUATION METRICS
R2: 0.70485
MAE: 0.53273
MSE: 0.52928
RMSE: 0.72752

Random Forest



EVALUATION METRICS
R2: 0.76338
MAE: 0.41071
MSE: 0.42433
RMSE: 0.65141

Conclusion

When considering steps a country would want to take regarding improving life expectancy, based on these models a country would be wise to focus on increasing health expenditure for each of its citizens (‘*healthexppercapita*’) and lowering the death rates from HIV or aids (‘*HIV.AIDS*’). According to the Random Forest Model, increasing the average number of years of schooling (‘*Schooling*’) and decreasing the rates of tuberculosis (‘*tuberculosis*’) are also significant factors, and according to the XGB model improving water access (‘*wateraccess*’) is also important in increasing life expectancy. Both these models had a high R-squared score of roughly 0.93, so these are fairly trustworthy models.

Regarding Fertility rates, a country looking to lower the rate of births would want to improve water access (‘*wateraccess*’) and increase the average number of years of schooling (‘*Schooling*’). The R-squared score is 0.76 for the Random Forest and 0.70 for the Ridge CV model, so while these models are okay they are not as reliable as the life expectancy models.

Random Forrest proved to be the best model for this dataset, scoring the highest R-squared scores and lowest MSE for both Fertility Rate and Life Expectancy outputs. Due to the Heteroscedasticity of this dataset, the results can be called into question, however given the high R-squared scores on the life expectancy models I would consider these to be accurate and useful models overall.

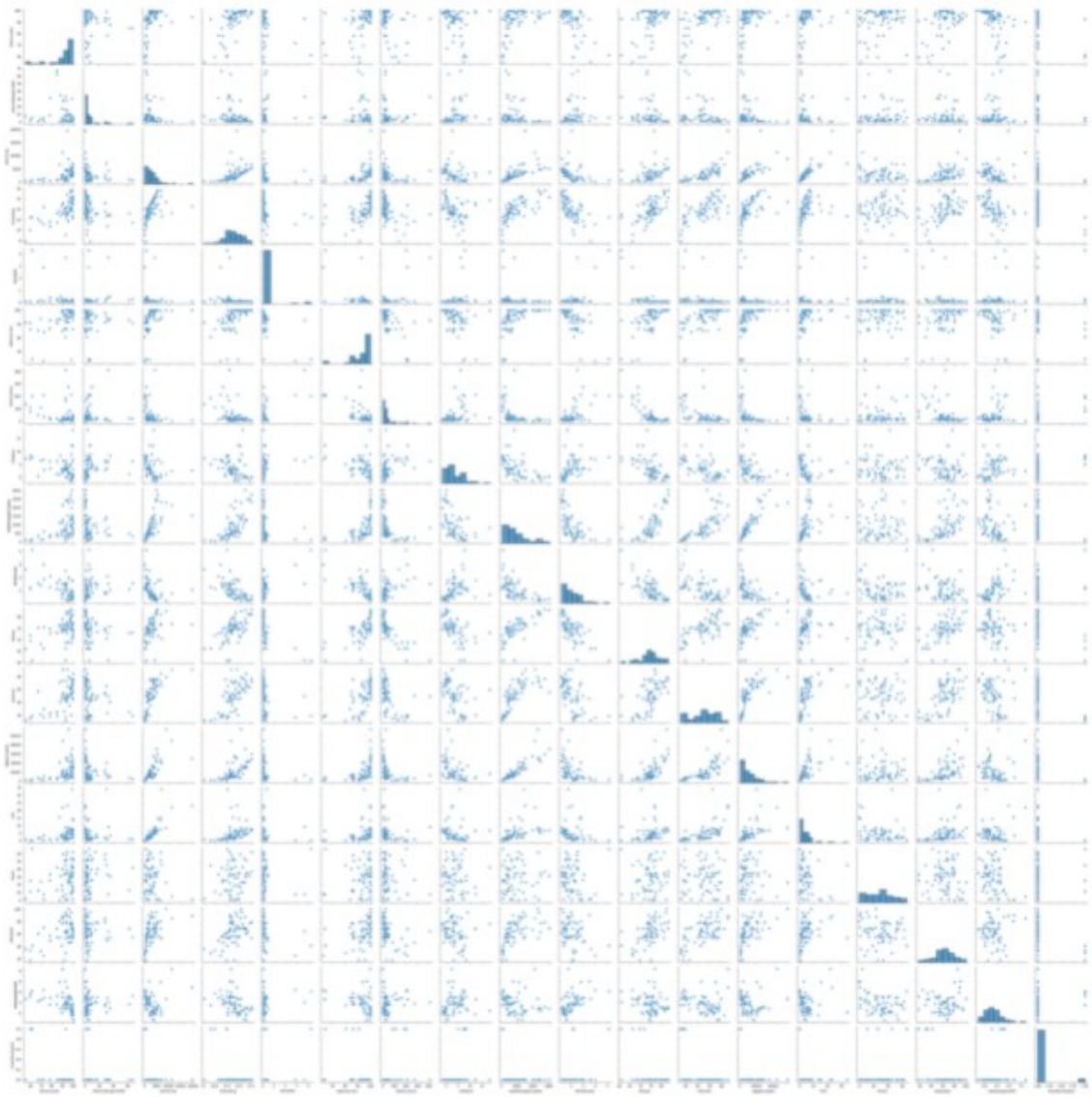
Lastly, it is worth noting that this dataset is not nearly large enough to reap the full benefits of regression analysis. While the evaluation metrics were decent, it is highly likely that another year of data or several more years would have a large impact on feature importance and coefficients, so this should be considered when evaluating these models.

Appendix:

Datasset 1 Pairplot

```
sns.pairplot(life_df_1)
```

```
<seaborn.axisgrid.PairGrid at 0x7f6ae50beb20>
```



Datasset 2 Pairplot

```
sns.pairplot(life_df_2)
```

```
<seaborn.axisgrid.PairGrid at 0x7f6ae56a9c70>
```