Predicting Life Expectancy - Springboard Capstone 2 by Sarah Carrier

Data proves that life expectancy has been increasing across the globe in both developing and developed countries. However, there is still room for prolonging human life spans, and maintaining healthy life expectancy should remain at the forefront of concern, particularly for countries deemed “developing” where life expectancy trails that of developed nations.

The World Health Organization tracks life expectancy by country as well as potentially correlating factors such as alcohol consumption, GDP, and immunization rates. The dataset tracks variables such as life expectancy over the course of 15 years, 2000-2015.

Factors like immunization rates probably influence life expectancy. Can the potentially correlating features be used to predict life expectancy for a country? If so, international leaders can then focus on those factors in order to boost life expectancy.

1. Data

WHO data about life expectancy is available via Kaggle.

2. Data Wrangling and Cleaning

Notebook:
https://github.com/atomlattice/Springboard-Capstone-2/blob/04e2bc3743707be6a2b75f56cabf95dc6f2a1c3c/data_wrangling_WHO_life_expectancy.ipynb

Summary:

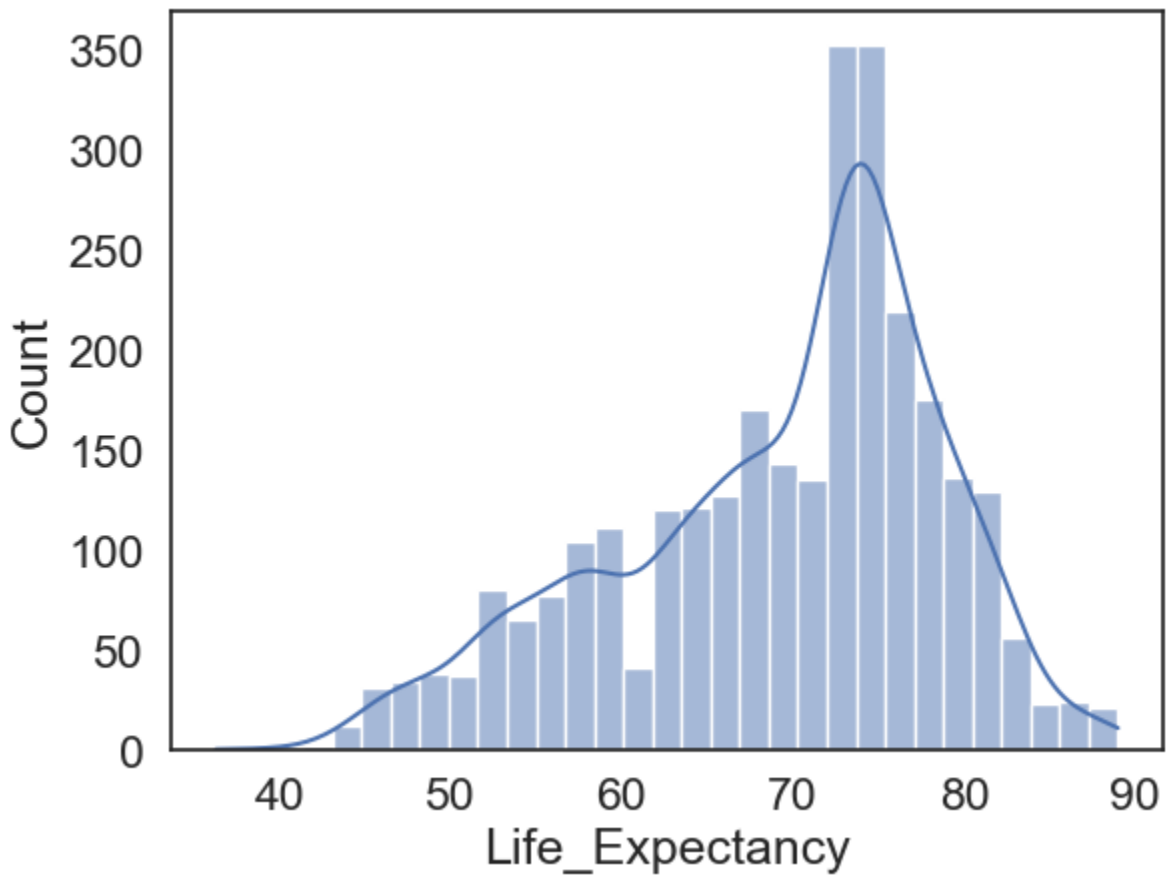
1. In total, 2938 rows and 22 features.
2. There were a lot of NaNs in the dataset and in fields where it was imperative that the values be generated to work with the data. To correct for this, I employed backwards and forwards interpolation to create values where needed.
3. Concerning 0s: many of the values needed to be 0, so I did not commit to a blanket replacement of all zeros.
4. There were also issues with spaces in the column names - I stripped the spaces at the end.

It became apparent as well that there were some problems with the dataset beyond NaNs. Some of the values in the original dataset were incorrect and could only be corrected by finding the original dataset from WHO. I attempted to do this but was unsuccessful - the origins of the dataset on Kaggle is therefore unclear. But wasn't to the extent that the analysis and modeling couldn't proceed, but would have implications for further work.

3. Exploratory Data Analysis

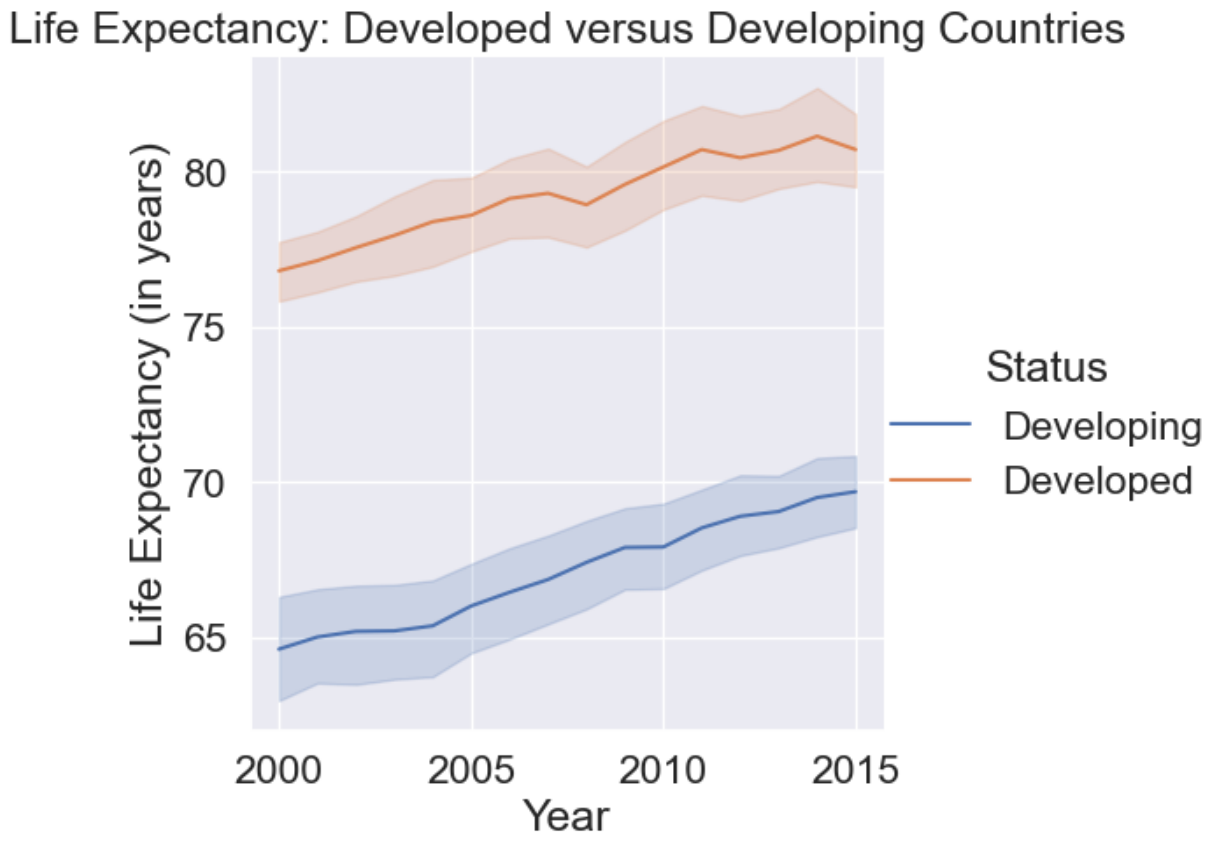
Notebook:https://colab.research.google.com/drive/1qJYCKc1dyuGLYUtaTyb7WxBw07qIUyV-?usp=sharing

A look at LIFE EXPECTANCY - the dependent variable



This visualization shows both the range and the mean of life expectancy: Range 45-90 years, Average lifespan 69 years.

A look at STATUS:



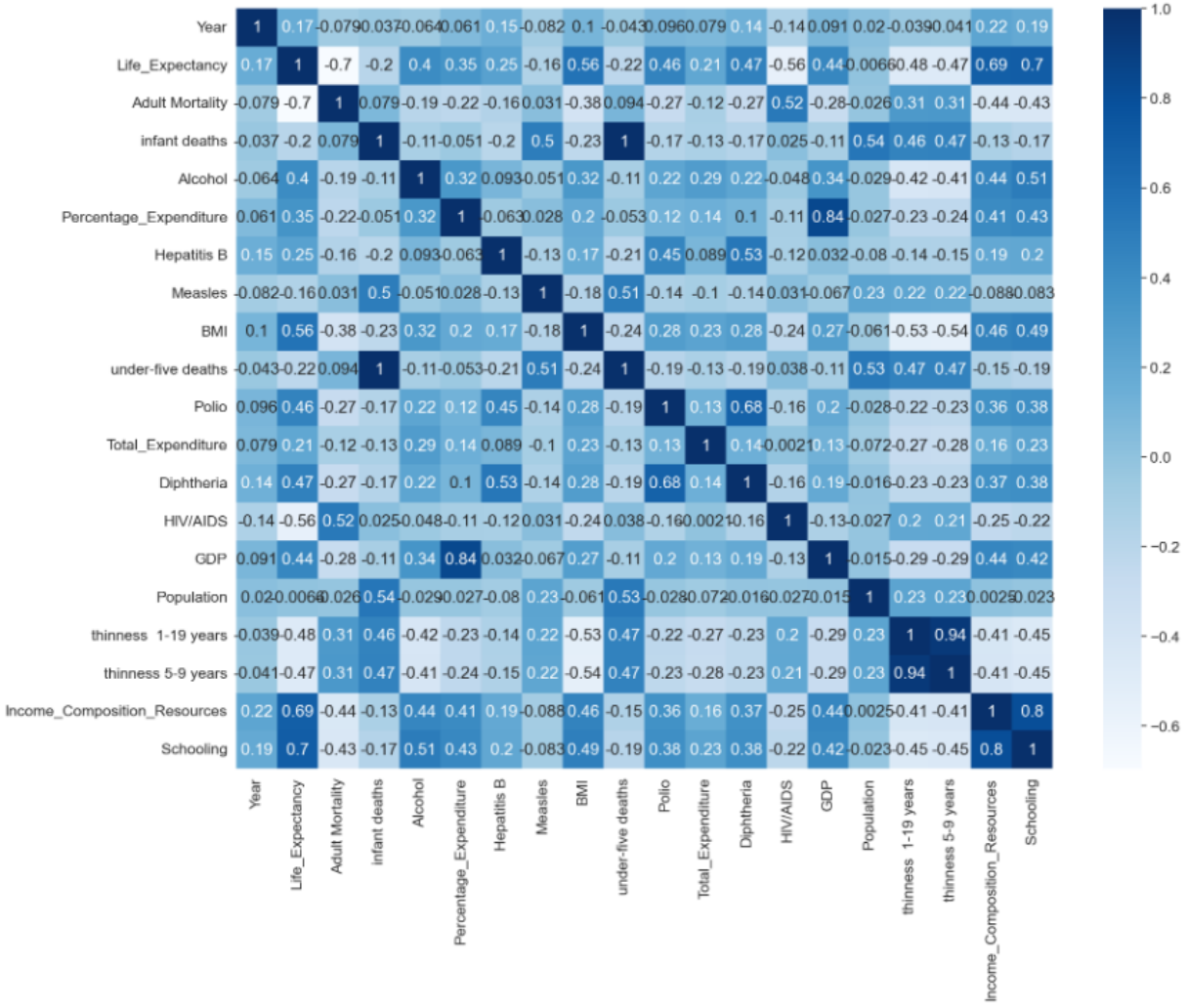
STATUS indicates whether a county is deemed “developing” or “developed”.

This is helpful to visualize because it shows

1. developing countries have lower life expectancy, generally, but
2. life expectancy has also been steadily increasing for both categories.

The category of developed or developing is of course also determined by other features and factors in this dataset - it is not a truly independent variable, rather another way for us to see how countries compare to one another.

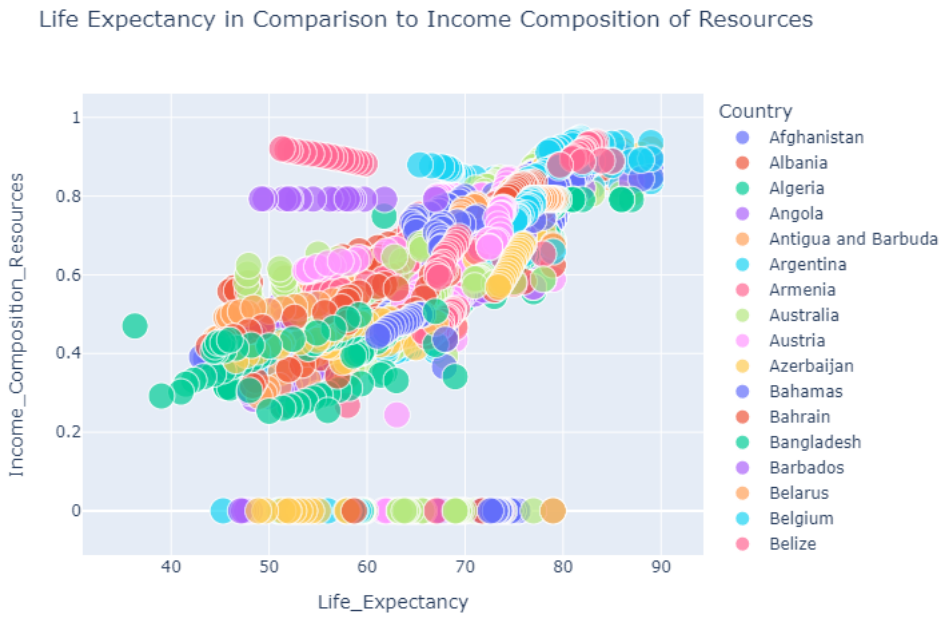
Correlations between features:



The heatmap shows that the most correlative features with life expectancy are **schooling** and **income composition of resources**.

Plotting features: Scatter plots revealed relationships between features and the dependent variable life expectancy. For example, plotting income composition of resources reveals a linear relationship with life expectancy.

Outliers turned out to be due to a problem with the original dataset itself.



4. Modeling and Machine Learning

Notebook:
https://github.com/atomlattice/Springboard-Capstone-2/blob/e0bcf699ae9a3ed03bf39d1acb51f7dc2f37ecc2/Modeling.ipynb

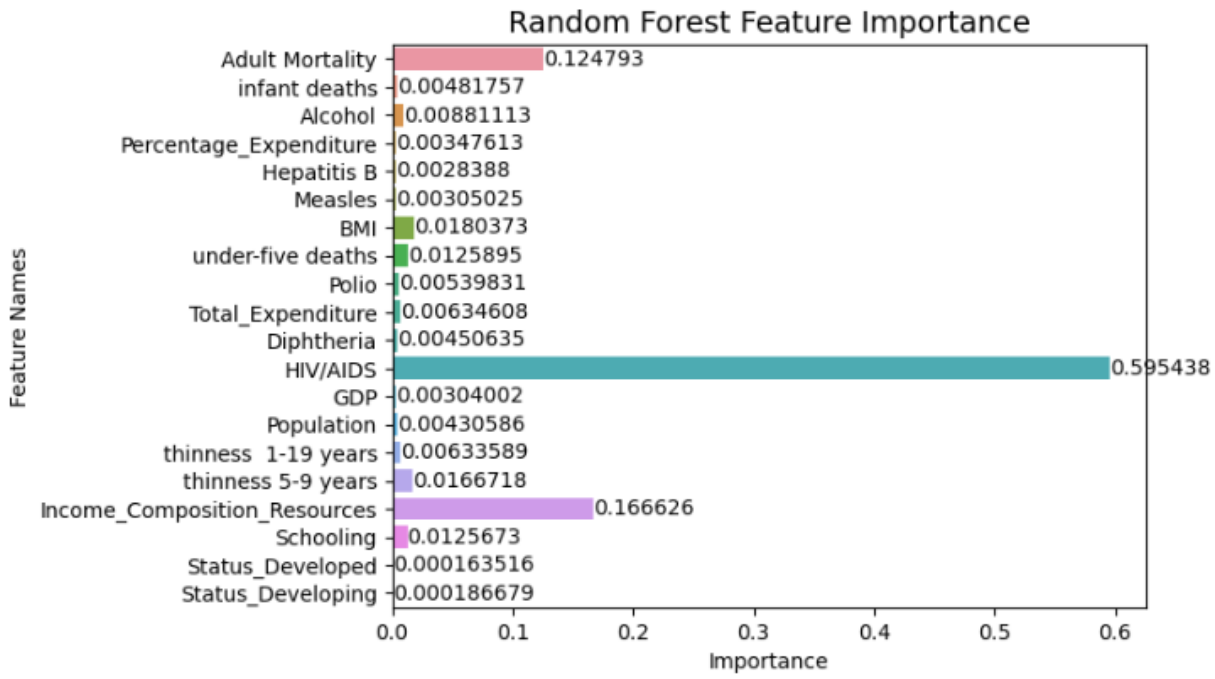
For this regression problem I chose to test the following models: Linear Regression, K Nearest Neighbors, Decision Tree, Random Forest, and Gradient Boosting.

I compared model performance based on R2 scores, Mean Squared Error

- Linear Regression score: 0.8085813264421107
- KNN score: 0.8696132136875967
- Decision Tree score: 0.9178792569811887
- Random Forest score: 0.9633052706154228
- Gradient Boosting score: 0.963120771276268

Result: **Random Forest** performed the best, although all the models performed *remarkably* well. And did not require tuning.

Feature importance with Random Forest:



Feature importance also shows income composition of resources. For further work, the correlation of HIV/AIDS should be examined. Like the variable STATUS, adult mortality is influenced by other features in this dataset, so it is not entirely independent.

8. Determination and Conclusions

Random Forest feature importance confirmed one of the feature correlations that became apparent during the EDA step: **income composition of resources (ICR)**

What is INCOME COMPOSITION OF RESOURCES? The extent to which the income composition in capital and labor income is distributed across the income distribution. High levels of income composition inequality are associated with class-fragmented societies, whereas low levels are typical of multiple-sources-of-income societies.

Income composition of resources is scored 0-1, with more equal societies have a score close to one.

Returning to some of the insights extracted during the EDA process: The country with the highest ICR in the dataset is NORWAY. Its average life expectancy across 15 years represented in the dataset is 82 – this is much higher than the average across the dataset of 69 years.

It is recommended that the first immediate step for increasing a country's life expectancy is to DECREASE inequality in income composition of resources.
The country of NORWAY should be looked at as it has a very high life expectancy and the highest ICR in the entire dataset of almost 1 (0.984). What is Norway doing to ensure an equitable income composition of resources? This approach should be modeled.

9. Future Work

Some of the values of the dataset appeared from the outset to potentially be wrong. This did not impede this project, but errors should be confirmed and corrected for future work.

Two features that indicated potential correlations but were not consistent across analysis or modeling should be looked into to determine if more scaling or cleaning could yield further insights:

- SCHOOLING – this was the feature of highest correlation according to the heat map
    - Schooling in this dataset is the average number of years a country's citizens spend in school.
- HIV/AIDS – this feature was by far scored the most important by the Random Forest model. In previous analysis, this feature did not seem to be important.
    - In this dataset, HIV/AIDS is "Deaths per 1 000 live births HIV/AIDS (0-4 years)"