Joseph Safer-Bakal

Data 1030

Professor Andras Zsom

12/10/2022

Link to Github Repo

## Introduction

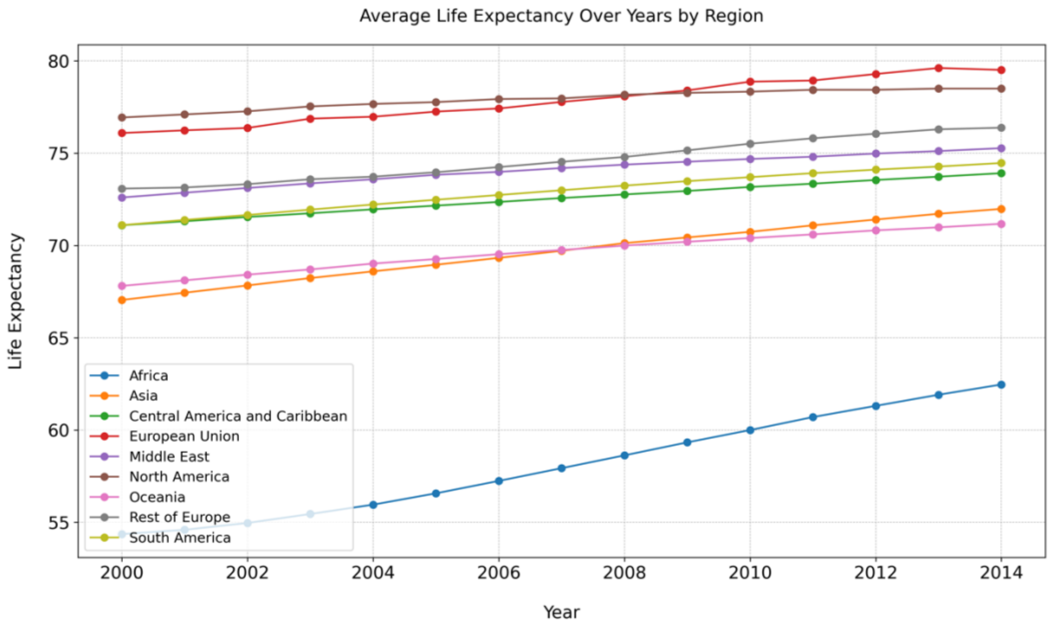
For my machine learning project, I sought to use a WHO dataset about global life expectancy to create machine learning models which predict a country's life expectancy for a given year based on a number of features, like BMI, disease rates, status as developing vs developed, and malnourishment from the preceding year. Through exploring this question, we may identify predictors for life expectancy and create a model that can help predict life expectancy from other variables, which would help policymakers globally in determining what policies might improve life expectancy in their respective countries in upcoming years.

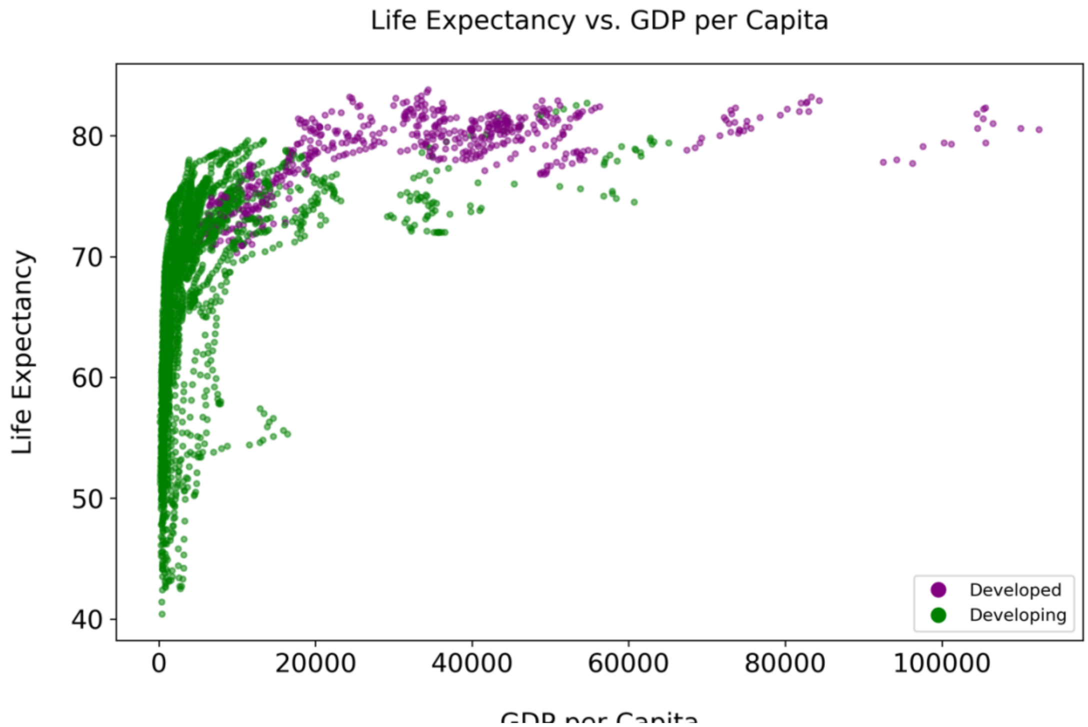
I sourced the data from Kaggle. The original WHO dataset was from the Global Health Organization (GHO) data repository and included data on 193 countries from 2000-2015. Kaggle user 'Lasha' contributed some initial preprocessing. They excluded countries with more than 4 columns of missing values and filled in remaining missing values with the closest 3-year average (for countries with only one year missing) or with the regional average (for countries with all years missing). The final dataset I accessed included 179 countries and 21 different metrics.

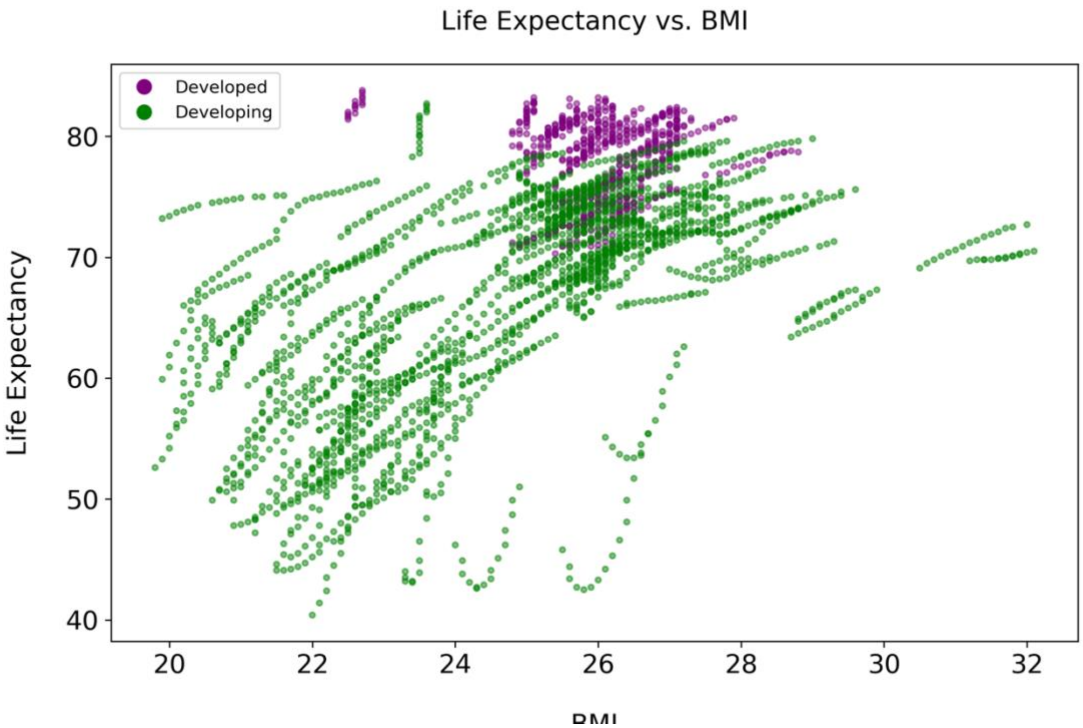
The target variable is life expectancy. In preprocessing, I adjusted the dataset to shift life expectancy up by one year so that the target variable would be the life expectancy for the following year.

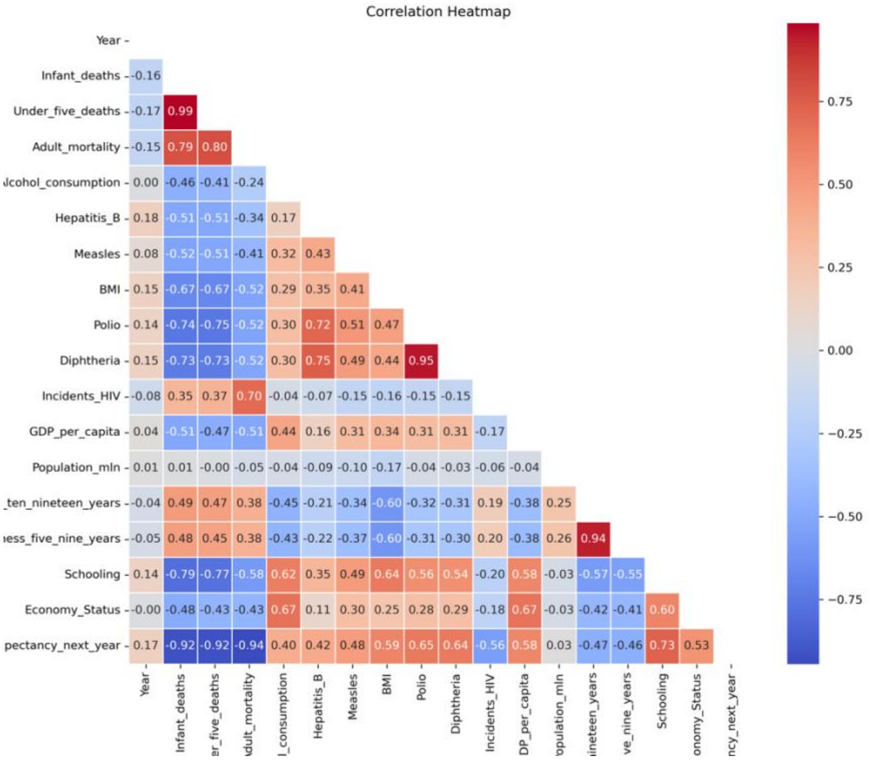
## EDA

In conducting EDA, I explored what features correlated most clearly with the target variable and how the variables correlated with each other:


Average Life Expectancy Over Years by Region
Life Expectancy
80
75
70
65
60
55
Year
2000
2002
2004
2006
2008
2010
2012
2014
Africa
Asia
Central America and Caribbean
European Union
Middle East
North America
Oceania
Rest of Europe
South America


Life Expectancy vs. GDP per Capita
Life Expectancy
80
70
60
50
40
0
20000
40000
60000
80000
100000
GDP per Capita
Developed
Developing


Life Expectancy vs. BMI
Developed
Developing
Life Expectancy
40
50
60
70
80
20
22
24
26
28
30
32
BMI


Correlation Heatmap
Year
Infant_deaths -0.16
Under_five_deaths -0.17 0.99
Adult_mortality -0.15 0.79 0.80
lcohol_consumption 0.00 -0.46 -0.41 -0.24
Hepatitis_B 0.18 -0.51 -0.51 -0.34 0.17
Measles 0.08 -0.52 -0.51 -0.41 0.32 0.43
BMI 0.15 -0.67 -0.67 -0.52 0.29 0.35 0.41
Polio 0.14 -0.74 -0.75 -0.52 0.30 0.72 0.51 0.47
Diphtheria 0.15 -0.73 -0.73 -0.52 0.30 0.75 0.49 0.44 0.95
Incidents_HIV -0.08 0.35 0.37 0.70 -0.04 -0.07 -0.15 -0.16 -0.15 -0.15
GDP_per_capita 0.04 -0.51 -0.47 -0.51 0.44 0.16 0.31 0.34 0.31 0.31 -0.17
Population_mln 0.01 0.01 -0.00 -0.05 -0.04 -0.09 -0.10 -0.17 -0.04 -0.03 -0.06 -0.04
_ten_nineteen_years -0.04 0.49 0.47 0.38 -0.45 -0.21 -0.34 -0.60 -0.32 -0.31 0.19 -0.38 0.25
ıess_five_nine_years -0.05 0.48 0.45 0.38 -0.43 -0.22 -0.37 -0.60 -0.31 -0.30 0.20 -0.38 0.26 0.94
Schooling 0.14 -0.79 -0.77 -0.58 0.62 0.35 0.49 0.64 0.56 0.54 -0.20 0.58 -0.03 -0.57 -0.55
Economy_Status -0.00 -0.48 -0.43 -0.43 0.67 0.11 0.30 0.25 0.28 0.29 -0.18 0.67 -0.03 -0.42 -0.41 0.60
pectancy_next_year 0.17 -0.92 -0.92 -0.94 0.40 0.42 0.48 0.59 0.65 0.64 -0.56 0.58 0.03 -0.47 -0.46 0.73 0.53
Year
Infant_deaths
er_five_deaths
dult_mortality
l_consumption
Hepatitis_B
Measles
BMI
Polio
Diphtheria
Incidents_HIV
DP_per_capita
opulation_mln
ıineteen_years
ve_nine_years
Schooling
onomy_Status
ncy_next_year
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

I found that some features correlated quite highly with one another, like Under_five_deaths and Infant_Deaths, as well as Thinness_five_nine_years and Thinness_ten_nineteen_years (measures of childhood and adolescent malnourishment). This makes sense, as much of the under-five mortality rate would be captured in the infant mortality rate. In order to make my models more accurate, I only included one feature out of two if they had greater than 90% correlation.

**Methods**

I used a time-based splitting strategy to divide my dataset into training, validation, and test sets. I used this approach because I am dealing with time-series data, and I my models are evaluated based on its ability to predict future life expectancy from the training data. This mimics real-world applications—policymakers trying to improve future life expectancy would only have access to data from the latest year. The training set comprises all data up to the year 2011. The validation set consists of data from the years 2012 to 2013. The test set includes data from 2014 onward. This sequential time-based split aimed to avoid having the model inadvertently learn from future data (which it would not have access to in a real-world scenario.

In my preprocessing, continuous features were standardized to have a mean of zero and a standard deviation of one. This standardization, however, was implemented later, within the machine learning pipeline so that only models that required standardization, like linear regression, would utilize it. For categorical features, I used onehotencoding. After preprocessing, the feature sets were concatenated to form the final datasets used for model training and evaluation.

My machine learning pipeline began with a preprocessing step that standardized continuous features, if specified, as I discussed in the splitting strategy section. I then created a machine learning algorithm in which I later plugged in my four machine learning models. Creating a single pipeline with preprocessing and an estimator built in ensured that the proper steps in training the models were executed sequentially, preventing data leakage. In order to tune the hyperparameters of my models, I used

GridSearchCV alongside a TimeSeriesSplit, which divided the dataset into five sequential folds, respecting the chronological order of the data. This was important because my data involves time-series analysis. I used GridSearchCV to conduct a search over a parameter grid (specified later when I actually trained the individual models), evaluating the model's performance across all combinations using the Root Mean Squared Error (RMSE) as the scoring function. I used RMSE as an evaluation metric because provides a clearer measure of the model's prediction error magnitude. I used Lasso Regression, Ridge Regression, Random Forest Regression, and XgBoost as my four machine learning models.

## Results

### Calculating Baseline Score

To calculate my baseline score, I took the last life expectancy from each country in the training set and imputed that as the predicted life expectancy for each country in the validation and training set.

The baseline RMSE was 0.9518383663234077

| | Baseline | Lasso Regression | Ridge Regression | Random Forest Regression | XGBoost |
|---|---|---|---|---|---|
| Train RMSE | N/A | 0.45884945144157846 | 0.39037262902275105 | 0.21490605626466214 | 0.16525851445216838 |
| Test RMSE | 0.9518383663234077 | 0.7119115550217102 | 0.7008691580419412 | 1.0643462237816292 | 0.8017392308390303 |
| Std Deviation | N/A | 0.08658572145292236 | 0.09372287914301412 | 0.31541554799303684 | 0.17506264727716703 |
| Parameters Tuned | N/A | Alpha | Alpha | Max_depth, max_features, n_estimators | N_estimators = 900, learning_rate, Max_depth |
| Best Parameters | N/A | Alpha = .001 | Alpha = .001 | Max_depth = 25, max_features = None, n_estimators = 40 | N_estimators, learning_rate = .3, Max_depth = 3 |
| Test RMSE Std Deviations Below Baseline | N/A | 2.77097432781857 | 2.6777795408793006 | -0.3441627736678806 | 1.1255458505732456 |

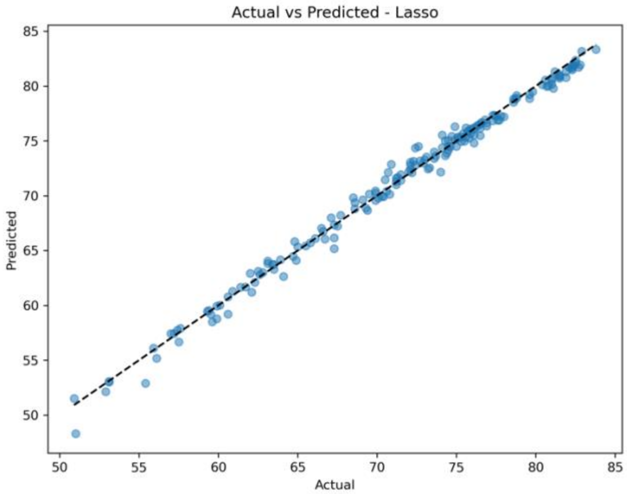
Most of the models outperformed the baseline RMSE. With the exception of the Random Forest Regression, all models had a lower RMSE on the test set, suggesting they are more accurate in predicting outcomes than simply predicting the most recent life expectance per country.

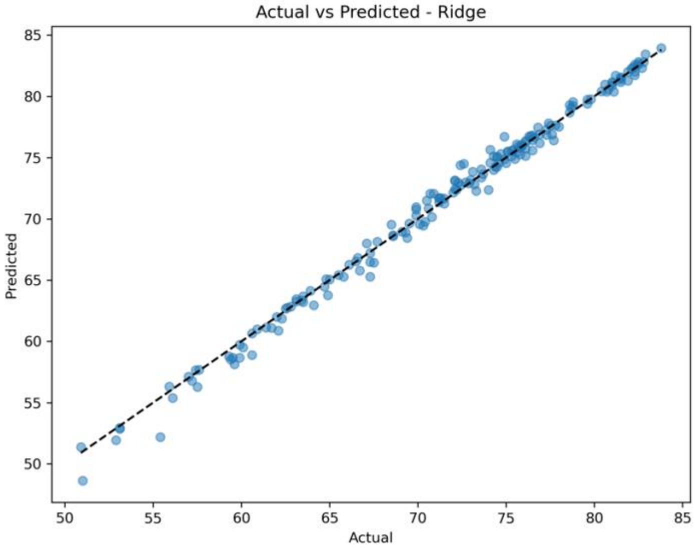
The Lasso and Ridge Regression models performed similarly, with test RMSEs of 0.7119 and 0.7009, respectively, and are about 2.77 and 2.68 standard deviations below the baseline—suggesting good performance relative to the baseline. The closeness of their RMSEs indicated that both models were effective for this dataset. The Ridge Regression had a slightly lower test RMSE, which may be because Ridge Regression handles multicollinearity between features better than Lasso Regression.

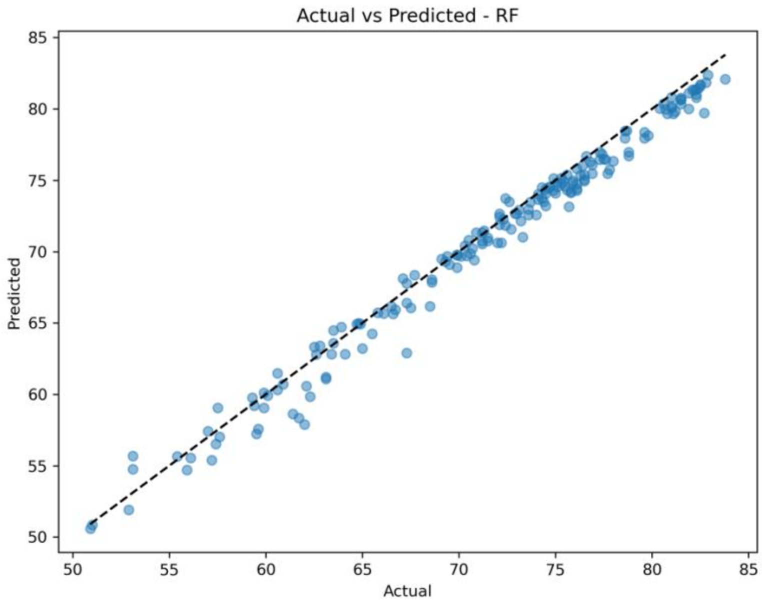
The Random Forest Regression had a test RMSE of 1.0643, which is 0.344 standard deviations above the baseline. The complexity of the Random Forest model might have contributed to it especially overfitting, as well as the fact that it proved too computationally intensive for me to tune more than the specified hyperparameters. The XGBoost model, with a test RMSE of 0.8017, performed better than the baseline; it's RMSE was 1.126 standard deviations below the baseline RMSE.

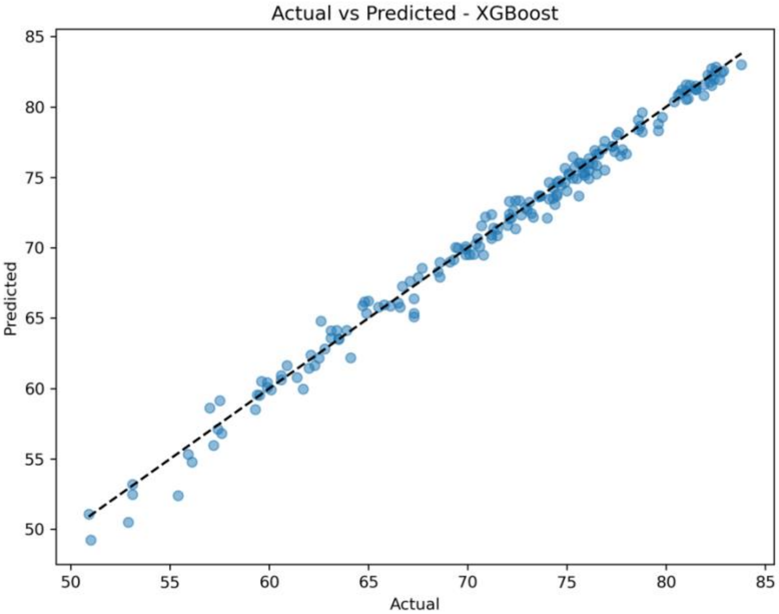
Overall, the models overfitted, as they all had lower training RMSEs relative to their test RMSEs. Regularization in Lasso and Ridge, and proper hyperparameter tuning in XGBoost, helped mitigate this to some extent, but the Random Forest model seems to have struggled more in this regard. I attempted to further reduce overfitting by further tuning hyperparameters and attempting to remove some features that might be causing overfitting but was ultimately unsuccessful in this regard.

The following graphs indicate that all models, with the exception of the RFR, performed reasonably well; the predicted values are quite close to the actual values, as represented by the black dotted line.

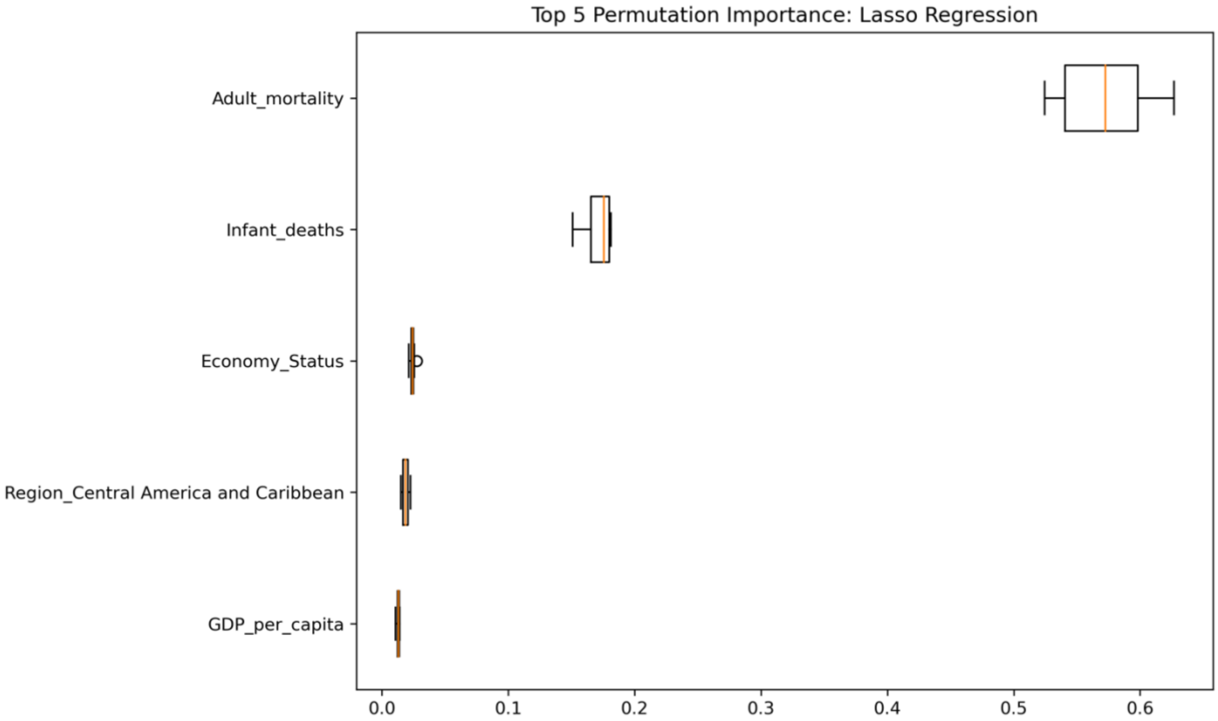

Actual vs Predicted - Lasso
85
80
75
70
65
60
55
50
Predicted
50
55
60
65
70
75
80
85
Actual

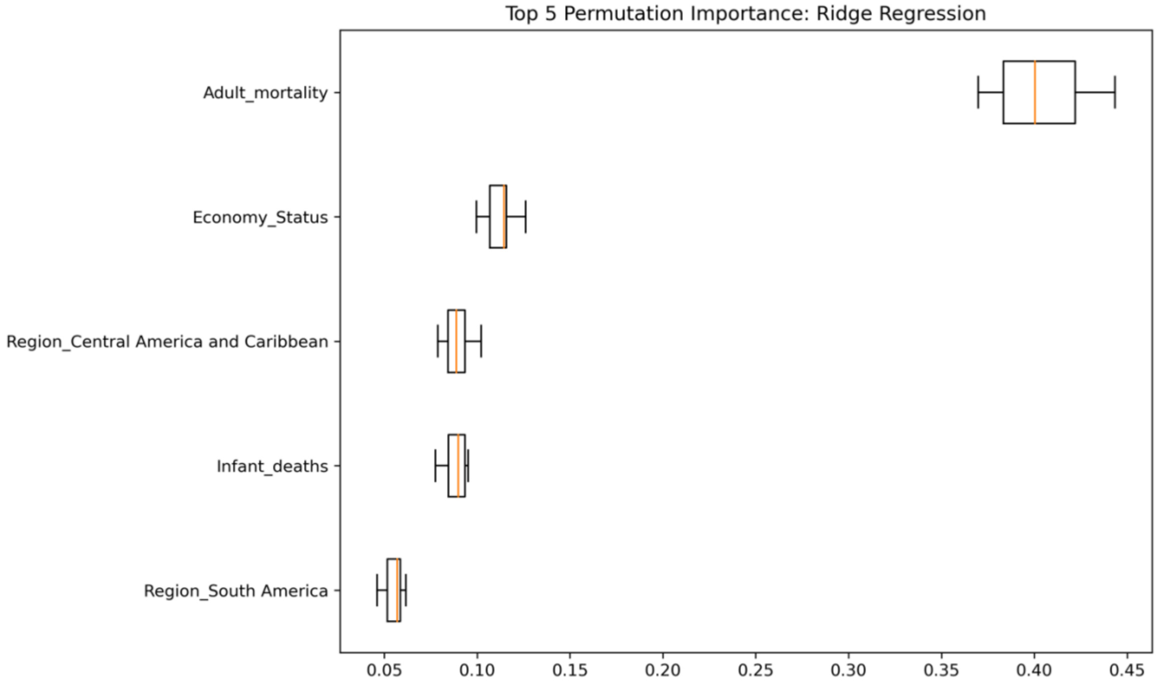

Actual vs Predicted - Ridge
85
80
75
70
65
60
55
50
Predicted
50
55
60
65
70
75
80
85
Actual

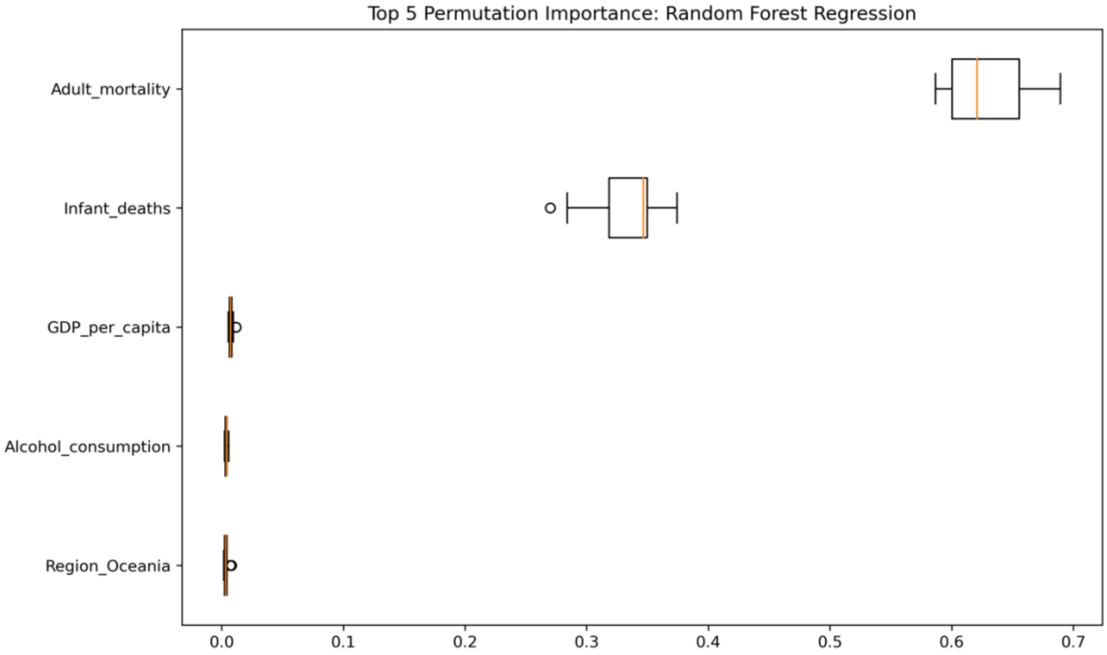

Actual vs Predicted - RF
Predicted
Actual
50
55
60
65
70
75
80
85
50
55
60
65
70
75
80
85

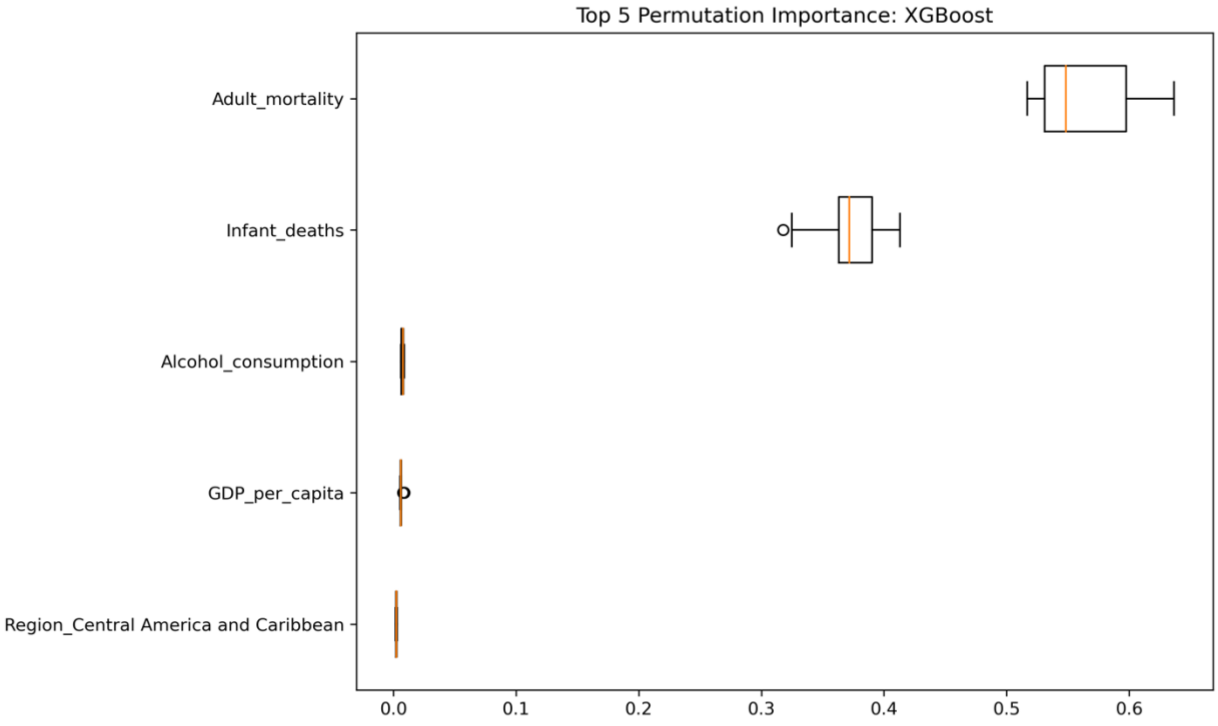

Actual vs Predicted - XGBoost
Predicted
Actual
50
55
60
65
70
75
80
85
50
55
60
65
70
75
80
85

In the following figures, I use permutation feature importance to show the top five features in the four models:




Top 5 Permutation Importance: Ridge Regression
Adult_mortality
Economy_Status
Region_Central America and Caribbean
Infant_deaths
Region_South America
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

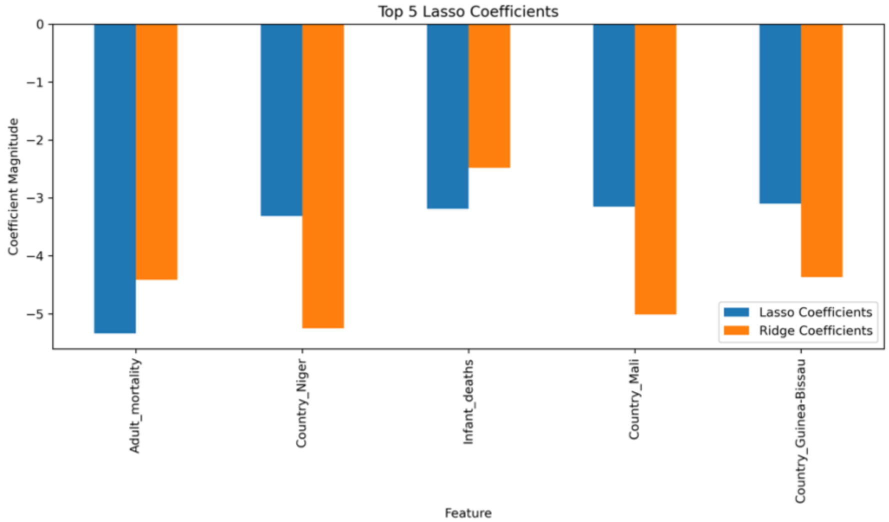

Top 5 Permutation Importance: Random Forest Regression
Adult_mortality
Infant_deaths
GDP_per_capita
Alcohol_consumption
Region_Oceania
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

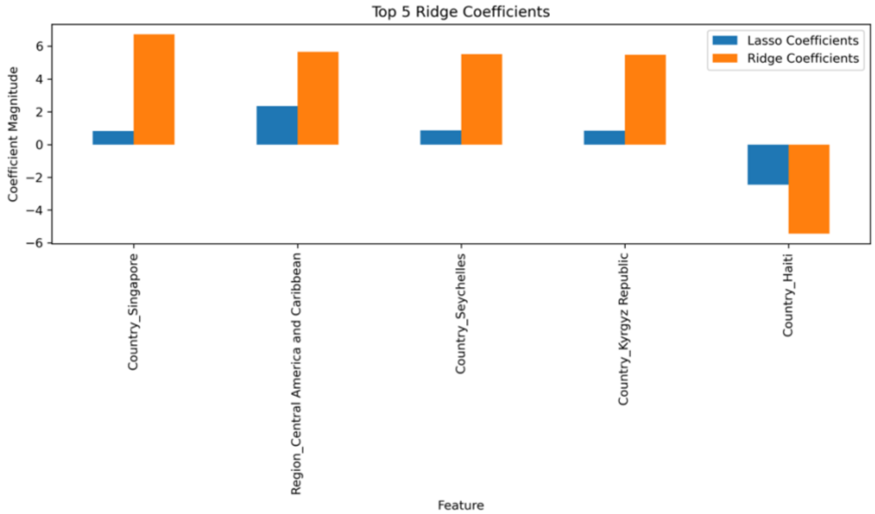


In the ridge regression model, “Adult_mortality,” “Infant_deaths,” and “Region_Central America and Caribbean” contributed significantly to the prediction. The regional feature suggests that the model is sensitive to geographic variations, which is not hugely helpful to local policymakers, as they of course cannot change the country they are in but may provide helpful info to NGOs seeking to target international aid by region. “Economy_Status” was also influential, suggesting that improved economic outcomes contribute to a higher life expectancy.

In the lasso regression model, “Adult_mortality” and “Infant_deaths” were also prominent, as well as “Region_Central America and Caribbean.” The premutation importance charts for the XGBoost and Random Forest Models also show “Adult_mortality” and “Infant_deaths” as significant across both models. This makes sense, as higher death rates should certainly indicate a lower a country’s life expectancy. This however is not terribly useful for policymakers because it doesn’t show cause of death, only higher death rates. “Alcohol_consumption” and “GDP_per_capita” also show up in both models but with varying degrees of importance.

The following graphs show the feature coefficients of the top five features for my lasso and ridge regressions, which represent the strength and direction of the relationship between each feature and the target variable after regularization has been applied.

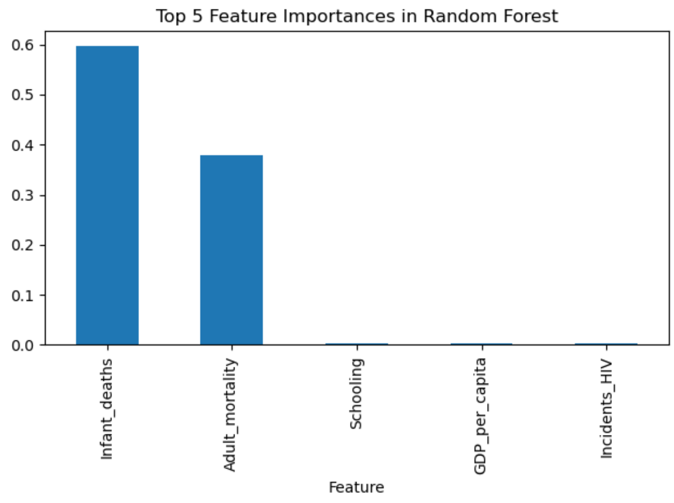


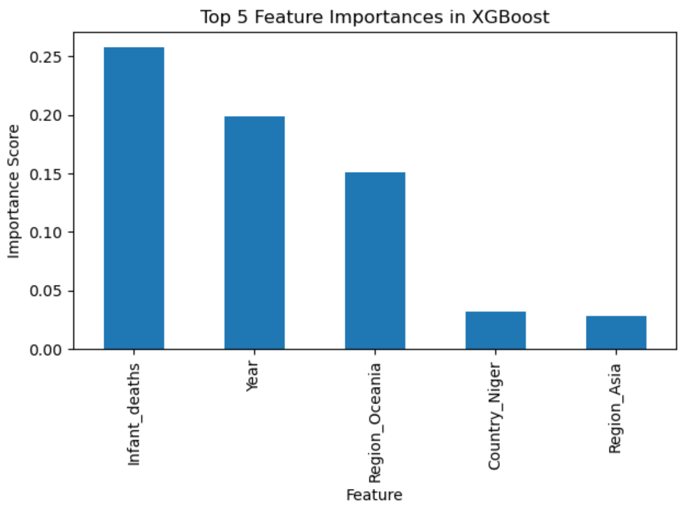


Each chart has the top five features for the specified model, as well as their contribution to the other model’s prediction.

The Lasso model chart shows that the coefficients of "Adult_mortality," "Infant_deaths," and features corresponding to specific countries like "Country_Mali," and "Country_Guinea-Bissau" were the largest. Since these features' impacts were negative, positive values for these features were associated with a decrease in next year's life expectancy. In the Ridge regression model, regional and national features had the most significant coefficients. Interestingly, however, some developing countries, like Seychelles, or a region with many developing countries, like South America and the Caribbean, contributed positively to the ridge regression's output. Country and regional coefficients seemed to have a larger role in the ridge regression; this difference is likely a result of the different ways in which lasso and ridge handle multicollinearity and feature selection.

The following graphs show the feature coefficients of the top five features for my Random Forest Regressor and XGBoost Models:

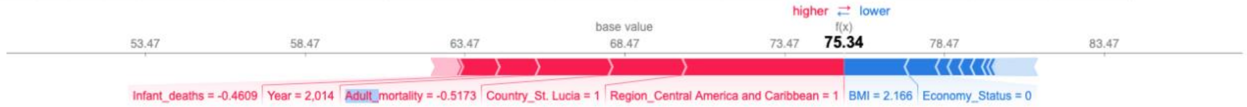




Confirming the results of the permutation importance graphs, in the Random Forest model, “Infant_deaths” and “Adult_mortality” had the greatest importance scores, followed by “Schooling,” “GDP_per_capita,” and “Incidents_HIV.” The latter three features seem to play a substantial role but to a lesser extent compared to the top two features.

In the XGBoost model, “Infant_deaths” again appeared as the most influential feature, followed by “Year” and “Region_Oceania.” “Year” also was a strong contributor—this may indicate that there were time-based trends that were relevant to the prediction.

This is a force plot visualizing the impact of each feature on a prediction made by my Ridge Regression Model:



This force plot visualizes the impact of each feature on the prediction made by the Ridge regression model for a specific instance—features pushing the prediction higher are shown in red, and those pushing the prediction lower are shown in blue. The base value (68.47) represents the average prediction over the dataset. The final prediction is the bolded number (75.34) above the plot and represents one prediction of the model after accounting for each feature's impact.

“Infant_deaths” and “Adult_mortality” have a negative relationship with the prediction but are pushing the prediction higher than the base value in this case. Despite the general trend of these features decreasing the prediction, the actual values of infant deaths and adult mortality in this case were low enough to not lower the prediction compared to the base value.

“BMI” pushed the prediction in a positive direction. A higher 'BMI' is likely an indicator of better nutrition status, and this relationship with improved life expectancy suggests one avenue for policymakers seeking to improve life expectancy.

'Economy_Status' had no impact in this instance. This could suggest that for this specific instance, the country’s economic status might not be a differentiating factor, or it is at its mean level, so it did not contribute to any deviation from the average prediction.

## Outlook

Overfitting was an issue common across all my models. While I corrected for this to some extent through hyperparameter tuning and some feature engineering, I ultimately was not successful in stopping overfitting. Some strategies to improve my modeling might be to implement other algorithms, like support vector machines, or explore more exhaustive hyperparameter tuning methods for the random forest model and XGBoost models, I would benefit from greater computing power—my computer was not able to handle an increased number of hyperparameters to tune, which I believe certainly impacted their predictive power.

I would also benefit from improved data. While the relationship between infant mortality and life expectancy gives some indication of strategies policymakers can take to improve life expectancy, it does not show what is driving high infant mortality rates. Likewise, adult mortality had a strong relationship with life expectancy, which is a fairly obvious relationship. If there was more granular data about what causes deaths both as infants and adults, the models would be greatly improved.

## References

Life Expectancy (WHO) Fixed. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated. Accessed 10 Dec. 2023.