Illinois Institute of Technology

Statistical Analysis on Factors Influencing Life Expectancy

Team Members:

Cheng Jiang(CJ) [33.33%]

Rahul Nair (RN) [33.33%]

Yurun Liu (YL) [33.33%]

Abstract

Life expectancy is a major issue especially in third-world countries where proper amenities are not available, in which economic factors play an important part. This paper is talking about building a model for life expectancy using factors such as social, economic, immunization, etc. We are interested in exploring how variables affect life expectancy. For this problem, we do exploratory data analysis. It provides a big picture and gives us an idea on how to build a model for it. We use several techniques such as multiple linear regression (MLR), ridge regression, Lasso regression, Decision tree, Random Forest and Adaptive Boosting for our life expectancy problem. We will decide which model is better based on the adjusted r square and root of mean square values.

Keywords: Life expectancy, models, multiple linear regression, Decision tree, ridge regression, Lasso regression, Random Forest, Adaptive Boosting.

Introduction

Lifespan is a topic that accompanies an individual's life. With the development of science and technology and the passage of time, more and more attention are paid to human lifespan. In US, we have an average of 2,839,205 people death per day.[1] The average life expectancy is 78.7 years for people in US.[1] According to the Global Health Observatory (GHO), it has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the developed nation in the past 30 years. If we can find demographic factors associated with life expectancy, we can expect certain people to have a certain length of lifespan.

For this project, we would like to explore people’s life expectancy, and the demographic factors associated with life expectancy using a model to predict people’s lifespan. For that we will be applying certain machine learning algorithms ranging from Multiple Linear regression, shrinkage methods like Lasso and Ridge, Decision Tree, Bagging (Random Forest) and Boosting (adaptive) to see how each of these techniques are efficient enough to give us an accurate result with least overfitting possible.

During the implementation of these techniques, we saw that some techniques were really good with prediction but had overfitting issues while for some it was the opposite. What we understood in general was that sometimes we might need to do a trade-off between overfitting and accuracy to get the model as much as accuracy as possible with minimum overfitting.

Problem Statement

There have been studies on factors that affect life expectancy like gender, ethnicity, etc.[4], but there hasn't been much study done on factors like immunization, which can matter a lot as there are countries where the awareness for immunization is less which may lead to lower life expectancy[5]. Other factors can also play a role like the GDP of that country, etc.

So the solution here is to do a study that puts focus on factors like immunization, social, economic and other health related factors and see if there is any correlation among these factors and predict what can be the life expectancy of an individual who is from a particular country with all the above mentioned factors provided.

DataSource:

For this kind of study, accuracy of information is a must. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data were collected from the United Nation website. The sample is globally representative. Among all categories of health-related factors, only those critical factors were chosen which are more representative. For example, we have immunization related factors , mortality factors, economic and social factors.

Methodology

As described before, the Global Health Observatory (GHO) data repository under the World Health Organization (WHO) has data from the year 2000-2015 for 193 countries. Variables included in this study were life expectancy in age, country status, some disease cases, GDP and so on. The life expectancy in age is the outcome and other variables are demographics of interest. These variables were obtained from the life expectancy files.

First, we do an exploratory data analysis (EDA) for the dataset using certain data visualization methods. We shall use boxplot to check outliers and use winsorizing and log transformation to clean the outliers. We use histograms to check if a factor has normal distribution or not. Also, we need to compare the developing and developed country's life expectancy. We use plots like pie chart to see the percentage of developing and developed countries in this dataset. In addition, we will build a correlation matrix for the predictor variables to select the correlations between the income variables. We will also plot line plots and scatter plots for the data and see the relation with life expectancy.

Then, we start to think about the models. First, we have to change the country, status and year to dummy variables since they are categorical. And then we recheck and clean the data again to make sure there are no other issues from data. We will remove variables which have multicollinearity issues. After that, we will split the data into 70% for training and 30% for testing the dataset. To see if all the predictors are life expectancy, we will include all variables into the multiple linear regression, if that's not the case we will apply shrinkage methods like ridge and lasso regression. We will also try other models like decision tree, random forest and adaptive boosting and check their respective adjusted R square and the root mean square error scores for the data. We will depend on the overfitting and scores to decide which model is better.

Results and Discussion

Data Cleaning:

Uniform Column Names:

We first renamed column names to have a more clear intuition and better manipulation as the format of column names are not uniform. For example, we changed “Life Expectancy” to “Life_Expectancy” and so on.

Dealing with Missing Values:

After checking the null values in our data, we find out there are 14 out of 22 variables containing missing values. Thus, performing data cleaning is important in this case. There are many columns with null values but the amount of null values is not large enough to drop the whole column. Considering the case of outliers, filling the missing values with variable mean might not be a good idea. We first tried to interpolate the values using the variable ‘Country’. However, it turned out that interpolation did not fill the missing values and we got the same result as before. The possible reason is that many countries have the first value as null and this interpolation did not fill the first null entry. Then we decided to fill the missing values using median yearwise. The missing values are all filled after this step.

Dealing with Outliers:

We first plot out the boxplot of all the continuous variables and find out that every variable has outliers in it except for BMI. (Figure 1)

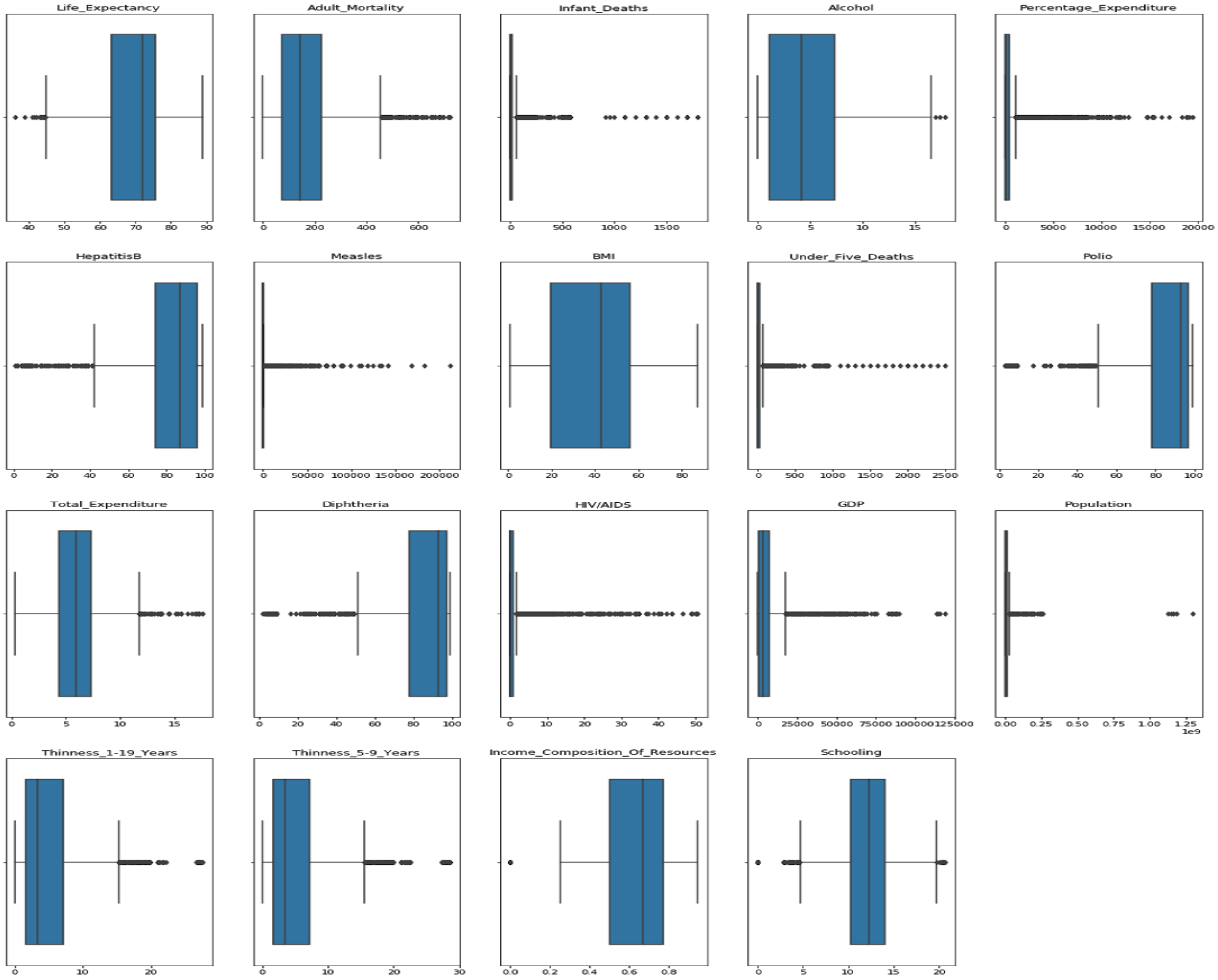


Figure 1: Boxplots of all continuous variables

In this case we decided to use winsorization technology to remove the outliers.We first wrote a function to count out the outliers numbers and outlier proportion for each variable and winsorize the variables according to the proportion of outliers we got in the last step. We then make a new data frame using the variables that are winsorized and filled up with missing values.

The boxplot of the new data frame shows that outliers are winsorized. (Figure 2)

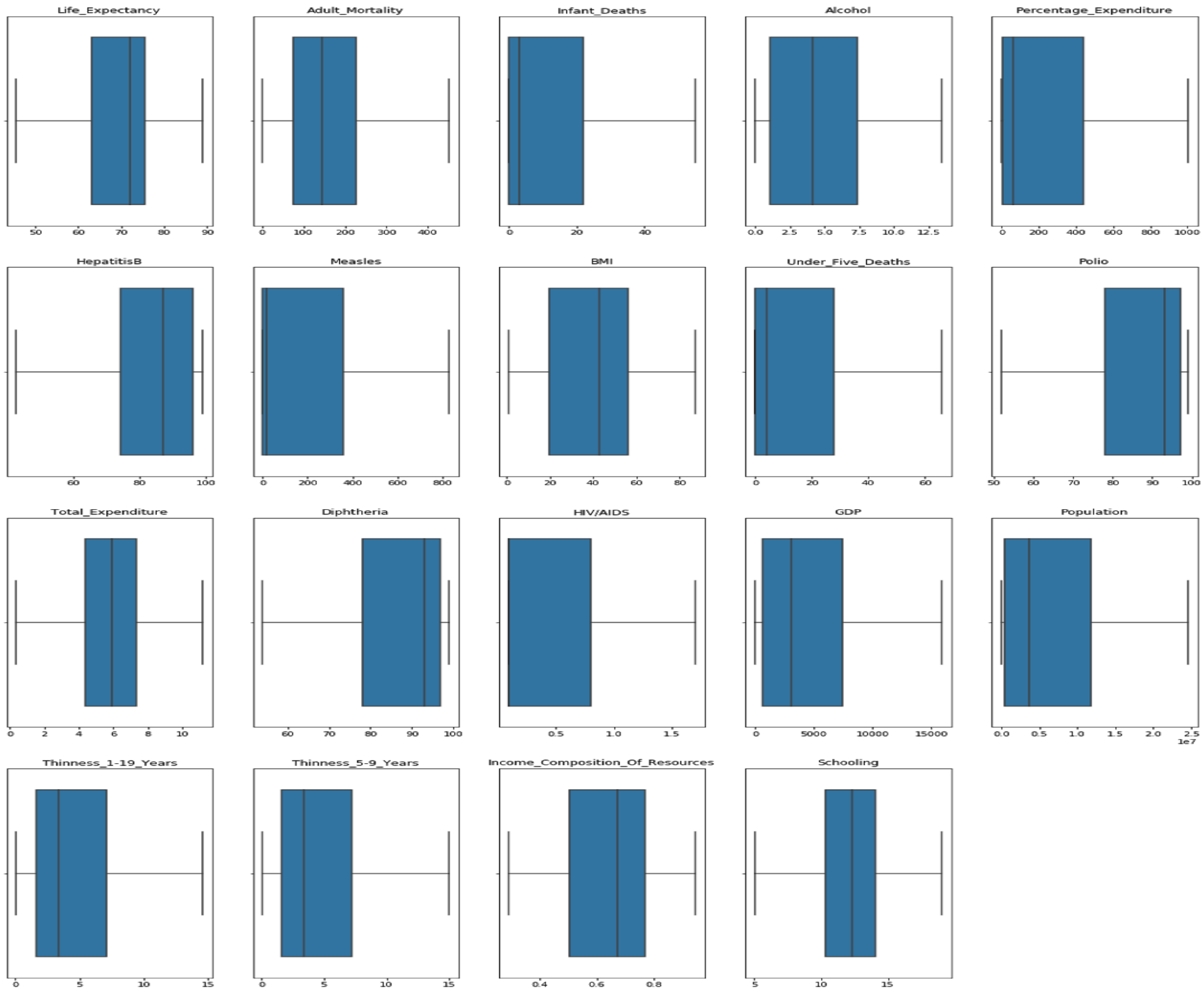


Figure 2: Boxplots of all continuous variables after winsorized

EDA:

A total of 193 countries were included in the current study. Most of the data is coming from developing countries and the life expectancy seems to be more in Developed countries (Figure 3). It does make sense as developed countries have better medical facilities to take care of citizen's health. We can see the life expectancy of developed countries is higher than developing countries' life expectancy (Figure 4).

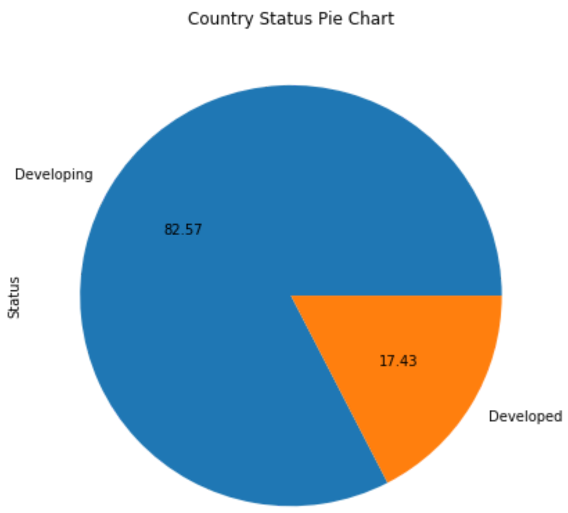


Figure 3: Country Status Pie Chart

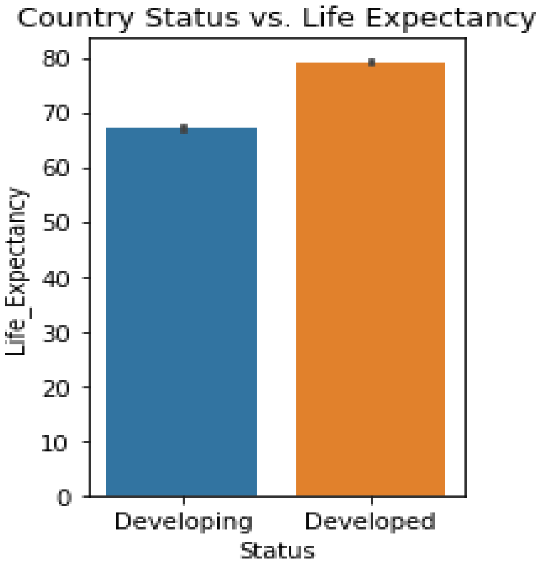
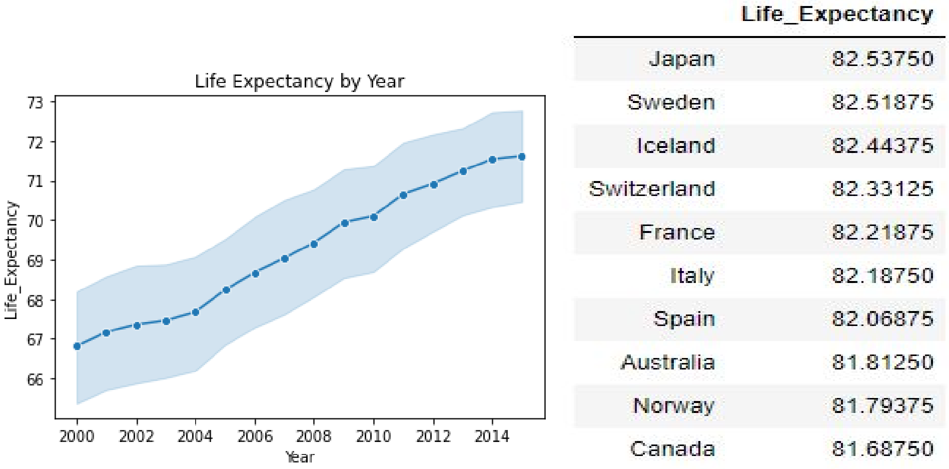


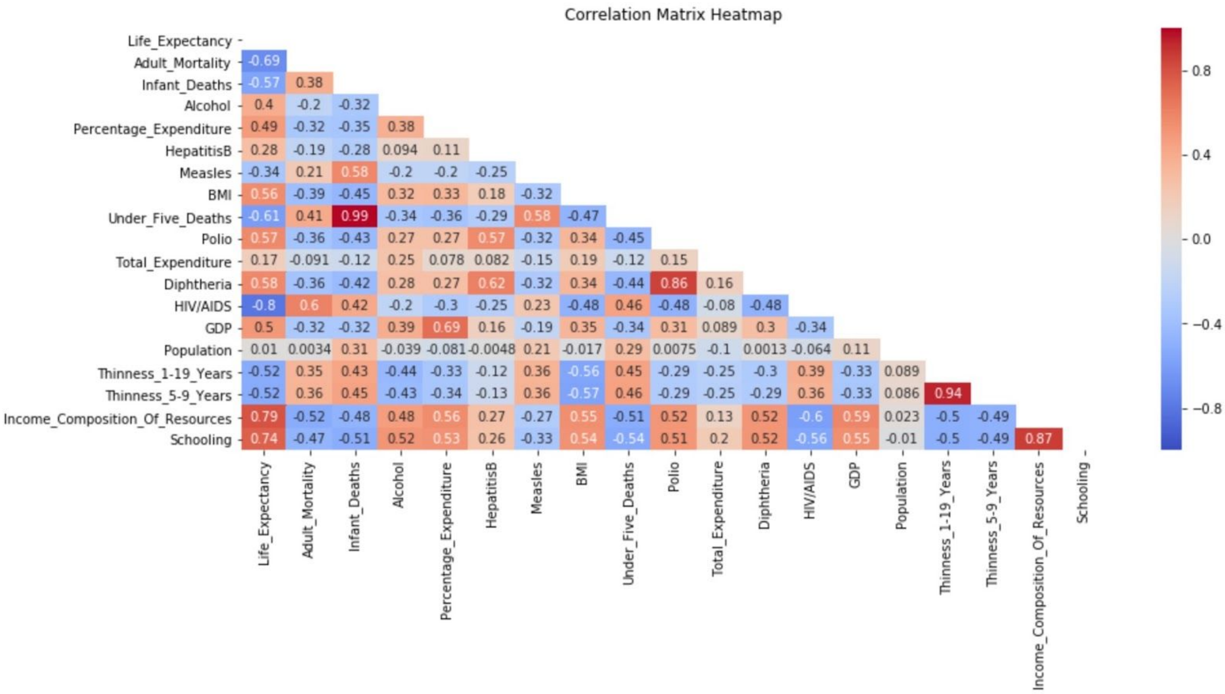
Figure 4 : country status VS life expectancy

The factors like income composition of resources gives us a good picture to show HIV/AIDS are highly correlated with the outcome life expectancy. Apart from that, we see some high correlations among the variables as well such as Diphtheria and Polio (0.88), and between Thinness_1-19_years and Thinness_5-9_years (0.97) which is very high (Figure 6). We also check about the life expectancy with years. We can easily see that life expectancy is on the rise with time (Figure 5). We list the top 10 countries for life expectancy and most of the countries are developed countries.



	Life_Expectancy
Japan	82.53750
Sweden	82.51875
Iceland	82.44375
Switzerland	82.33125
France	82.21875
Italy	82.18750
Spain	82.06875
Australia	81.81250
Norway	81.79375
Canada	81.68750

Figure 5: Life Expectancy by year



	Life_Expectancy	Adult_Mortality	Infant_Deaths	Alcohol	Percentage_Expenditure	HepatitisB	Measles	BMI	Under_Five_Deaths	Polio	Total_Expenditure	Diphtheria	HIV/AIDS	GDP	Population	Thinness_1-19_Years	Thinness_5-9_Years	Income_Composition_Of_Resources	Schooling
Life_Expectancy																			
Adult_Mortality	-0.69																		
Infant_Deaths	-0.57	0.38																	
Alcohol	0.4	-0.2	-0.32																
Percentage_Expenditure	0.49	-0.32	-0.35	0.38															
HepatitisB	0.28	-0.19	-0.28	0.094	0.11														
Measles	-0.34	0.21	0.58	-0.2	-0.2	-0.25													
BMI	0.56	-0.39	-0.45	0.32	0.33	0.18	-0.32												
Under_Five_Deaths	-0.61	0.41	0.99	-0.34	-0.36	-0.29	0.58	-0.47											
Polio	0.57	-0.36	-0.43	0.27	0.27	0.57	-0.32	0.34	-0.45										
Total_Expenditure	0.17	-0.091	-0.12	0.25	0.078	0.082	-0.15	0.19	-0.12	0.15									
Diphtheria	0.58	-0.36	-0.42	0.28	0.27	0.62	-0.32	0.34	-0.44	0.86	0.16								
HIV/AIDS	-0.8	0.6	0.42	-0.2	-0.3	-0.25	0.23	-0.48	0.46	-0.48	-0.08	-0.48							
GDP	0.5	-0.32	-0.32	0.39	0.69	0.16	-0.19	0.35	-0.34	0.31	0.089	0.3	-0.34						
Population	0.01	0.0034	0.31	-0.039	-0.081	-0.0048	0.21	-0.017	0.29	0.0075	-0.1	0.0013	-0.064	0.11					
Thinness_1-19_Years	-0.52	0.35	0.43	-0.44	-0.33	-0.12	0.36	-0.56	0.45	-0.29	-0.25	-0.3	0.39	-0.33	0.089				
Thinness_5-9_Years	-0.52	0.36	0.45	-0.43	-0.34	-0.13	0.36	-0.57	0.46	-0.29	-0.25	-0.29	0.36	-0.33	0.086	0.94			
Income_Composition_Of_Resources	0.79	-0.52	-0.48	0.48	0.56	0.27	-0.27	0.55	-0.51	0.52	0.13	0.52	-0.6	0.59	0.023	-0.5	-0.49		
Schooling	0.74	-0.47	-0.51	0.52	0.53	0.26	-0.33	0.54	-0.54	0.51	0.2	0.52	-0.56	0.55	-0.01	-0.5	-0.49	0.87	

Figure 6: Correlation Matrix Heatmap

Modelling:

First, we split the dataset into 70% training dataset and 30% testing dataset. Then we implemented multiple linear regression model. We find that most of the predictors have significant p-values (See Appendix B). The coefficient value makes sense. For example, we can see the life expectancy with a variable year (See Appendix A). The coefficient is positive which means it is increasing by year. The adjusted r square is 0.95 which is pretty good. The F statistic is 224.8 that proves our model is significant in predicting the outcome. To check the residuals, we draw a residuals vs fits plot (Figure 7). The line is kind of horizontal which proves the linear regression model is appropriate for the data. Also, the model generated is good enough as the points are getting clustered towards the center. However, when we calculate evaluation metrics for testing data, we got root mean square error (RMSE) as 2.182494, r square 0.946722, and Adjusted R square 0.946783, but we got evaluation metrics for training data RMSE as 3.158340, r square 0.963511, and Adjusted R square 0.959225. The test RMSE is greater than training RMSE little bit which means the model is a little bit overfitting. So, we try to use shrinking methods like lasso regression and ridge regression and see if we can reduce that.

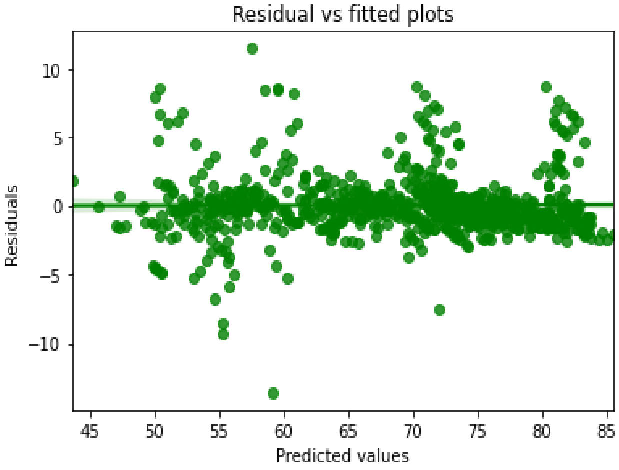


Figure 7: Residual vs fits

We used ridge regression and got the evaluation metrics for training. RMSE is 1.807396, r square is 0.963511, and Adjusted R square is 0.963529. The ridge regression evaluation metrics for testing RMSE is 2.184339, r square is 0.946632, and Adjusted R square is 0.946693. It still has an overfitting problem and the Adjusted R square gets worse.

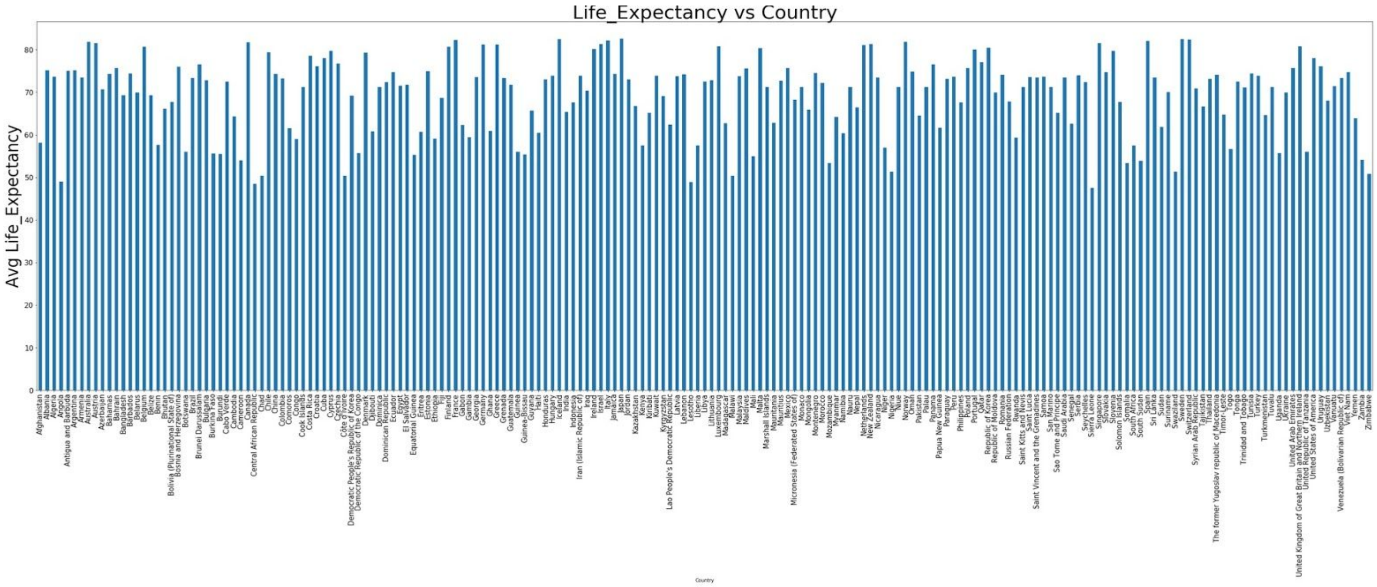
So, we try Lasso regression now. For training data, RMSE is 1.807747, r square is 0.963497, and Adjusted R square is 0.963514. The Lasso regression evaluation metrics for testing RMSE is 2.188782, r square is 0.946415, and Adjusted R square is 0.946415. The model is slightly better if we consider the overfitting issue, but not impressive. So, we are thinking about other models now. There are decision tree model and random forest model. When applied decision tree, for training data, RMSE is 1.953905, r square is 0.957355, and Adjusted R square is 0.957376. The decision tree evaluation metrics for testing RMSE is 2.684426, r square is 0.919399, and Adjusted R square is 0.919490. Since the decision tree was performing so badly, we thought of using a bagging technique like Random Forest which will run independent trees and try getting better results. And as expected, the results did improve. The evaluation metrics for training RMSE is 0.929401, r square is 0.990351, and Adjusted R square is 0.990356. The Random Forest evaluation metrics for testing RMSE is 1.979576, r square is 0.956169, and Adjusted R square is 0.956218. But there are signs of overfitting. Now let's try Boosting for which we will use Adaptive Boosting. The Adaptive Boosting evaluation metrics for training RMSE is 2.990664, r square is 0.900094, and Adjusted R square is 0.900142. The Adaptive Boosting evaluation metrics for testing RMSE is 3.168037, r square is 0.887741, and Adjusted R square is 0.887869. Now we see that the overfitting issue is slightly reduced but the overall accuracy is not impressive (See Appendix C). The best subset is a good way to improve our model but there are too many variables in this linear model. So, we did not use it to simplify our multiple linear regression.

Conclusion

In conclusion, based on the adjusted R square and RMSE values, we find that Random Forest came out to be the best model given that you are not considering overfitting issues. Because in terms of lowest overfitting, Adaptive Boosting performed the best having the least overfitting issue but the metric values are way less. Therefore we will have to do a trade-off between overfitting and accuracy. Considering that, we find Lasso Regression to be the best option since it has considerably least overfitting occurring and the evaluation metric values (both RMSE and Adjusted R square) are good. Choosing this gives us a good prediction power for predicting life expectancy

Appendix

Appendix A: Bar graph showing Life Expectancy across the countries.



Appendix B: Result and part of summary for multiple regression model

```
                              OLS Regression Results



     Dep. Variable:       Life_Expectancy    R-squared:                 0.964

     Model:                     OLS         Adj. R-squared:             0.959

     Method:              Least Squares      F-statistic:              224.8

     Date:             Thu, 03 Dec 2020     Prob (F-statistic):          0.00

     Time:                   12:01:14       Log-Likelihood:           -4134.3

     No. Observations:              2056      AIC:                    8703.

     Df Residuals:                1839      BIC:                    9924.

                        Df Model:                   216

                        Covariance Type:        nonrobust

                                          coef    std err      t       P>|t|    [0.025     0.975]
     ------------------------------------------------------------------------------------------------

     const                              46.3783    0.572    81.115    0.000    45.257    47.500

     Adult_Mortality                    -0.0010    0.001    -1.794    0.073    -0.002  9.72e-05

      Infant_Deaths                      0.1662    0.052     3.195    0.001     0.064     0.268

      Alcohol                           -0.0937    0.030    -3.081    0.002    -0.153    -0.034

      HepatitisB                        -0.0055    0.004    -1.431    0.153    -0.013     0.002

      Measles                           -0.0003    0.000    -1.309    0.191    -0.001     0.000
     BMI                              -0.0062    0.004   -1.725    0.085   -0.013    0.001

     Under_Five_Deaths                -0.2538    0.036   -7.049    0.000   -0.324   -0.183

     Polio                             0.0255    0.005    4.865    0.000    0.015    0.036

     Total_Expenditure                -0.1502    0.138   -1.091    0.276   -0.420    0.120

     HIV/AIDS                         -0.3464    0.276   -1.254    0.210   -0.888    0.195

  GDP                             -2.707e-05  1.32e-05   -2.056    0.040  -5.29e-05  -1.25e-06

  Population                       7.833e-09  7.87e-09    0.995    0.320   -7.6e-09   2.33e-08

     Thinness_1-19_Years              -0.0395    0.029   -1.375    0.169   -0.096    0.017

     Income_Composition_Of_Resources   2.1129    0.927    2.279    0.023    0.294    3.931
```

Appendix C: evaluation metrics for training data and testing data

Training data

	Model R_square	R_square	Root Mean_square	Adjusted R_square
0	Multiple Linear Regression	0.963511	3.158340	0.959225
1	Ridge Regression	0.963511	1.807396	0.963529
2	Lasso Regression	0.963497	1.807747	0.963514
3	Decision Tree	0.957355	1.953905	0.957376
4	Random Forest	0.990351	0.929401	0.990356

5	Adaptive Boosting	0.900094	2.990664	0.900142

Testing data

	Model R_square	R_square	Root Mean_square	Adjusted R_square
0	Multiple Linear Regression	0.947973	2.156721	0.948032
1	Ridge Regression	0.946632	2.184339	0.946693
2	Lasso Regression	0.946415	2.188782	0.946415
3	Decision Tree	0.919399	2.684426	0.919490
4	Random Forest	0.956169	1.979576	0.956218
5	Adaptive Boosting	0.887741	3.168037	0.887869

Bibliography

1. https://www.cdc.gov/nchs/fastats/deaths.htm
2. https://www.kaggle.com/kumarajarshi/life-expectancy-who
3. https://tungmphung.com/ensemble-bagging-random-forest-boosting-and-stacking/
4. https://www.simplyinsurance.com/average-us-life-expectancy-statistics/
5. https://countrydetail.com/top-10-countries-lowest-life-expectancy-rates-world/