1. **INTRODUCTION**
2. **Problem Statement:**

* To determine an efficient regression model to predict the effect of our predictor variables mentioned in the table on the response variable, that is, Life Expectancy.
* To run a simple machine learning algorithm on the dataset to avoid the problem of over-fitting.
* To run tests on the predictor variables in order to gauge which variables to use and to what extent do they affect our response variable.
* To determine Individual relationship between our predictors and response variable.
* To show any trends in the dataset- county-wise and continent-wise.
* To provide insight of how suicide rates have changed over the years.
* To predict what must change to increase a country's life expectancy.

1. **Data Sources:**



Kaggle - *https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who*

1. **Data Description:**

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
|  |  |  |
| Country | Name of countries | Character |
|  |  |  |
| Year | Year of measurement. | Numeric |
|  |  |  |
|  | Identifies the status of a country is |  |
| Status | Developing or developed. | Character |
|  |  |  |
|  | Identifies the average life span of an |  |
| Life Expectancy | individual | Numeric |
|  |  |  |
| Adult Mortality | Identifies the number of Adult Deaths | Numeric |
|  |  |  |
| Infant deaths | Identifies the number of infant deaths. | Numeric |
|  |  |  |
| Alcohol | Identifies Alcohol rate | Numeric |
|  |  |  |
|  | Identifies the amount of money a person |  |
| Percentage expenditure | spends of the total income. | Numeric |
|  |  |  |
|  | Identifies the number of people having a |  |
|  | serious liver infection caused by the hepatitis |  |
| Hepatitis B | B virus (HBV) | Numeric |
|  |  |  |
|  | Identifies the number of people having an |  |
|  | acute contagious disease that is caused by a |  |
| Measles | morbillivirus (species Measles morbillivirus) | Numeric |
|  |  |  |
| BMI | Identifies the Body Mass Index | Numeric |
|  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Identifies the number of children dying |  |  |
| Under-five deaths | between birth and exactly 5 years of age. | Numeric |  |
|  |  |  |  |
|  | Identifies the number of people having a |  |  |
|  | disabling and life-threatening disease by the |  |  |
| Polio | poliovirus. | Numeric |  |
|  |  |  |  |
|  | Identifies the total amount of money that is |  |  |
| Total expenditure | spent on a product in a given time period | Numeric |  |
|  |  |  |  |
|  | Identifies the number of people having |  |  |
|  | Diphtheria disease- a serious bacterial |  |  |
|  | infection that usually affects the mucous |  |  |
| Diphtheria | membranes of the nose and throat | Numeric |  |
|  |  |  |  |
|  | Identifies the rate of HIV/AID per 100,000 |  |  |
| HIV/AIDS | people. | Numeric |  |
|  |  |  |  |
| GDP | Identifies the Gross Domestic Product | Numeric |  |
|  |  |  |  |
| Population | Identifies the population | Numeric |  |
|  |  |  |  |
|  | Identifies the rate of thinness among people |  |  |
| Thinness 1-19 years | aged 10-19 | Numeric |  |
|  |  |  |  |
|  | Identifies the rate of thinness among people |  |  |
| Thinness 5-9 years | aged 5-9 | Numeric |  |
|  |  |  |  |
|  | Identifies the relative share of each income |  |  |
| Income composition of r | source or group of sources, expressed as a |  |  |
| percentage of the aggregate total income of |  |  |
| esources | Numeric |  |
| that group or area. |  |
|  |  |  |  |

|  |  |  |
| --- | --- | --- |
| Schooling | Identifies the school rating. | Numeric |
|  |  |  |
|  |  |  |
|  |  |  |

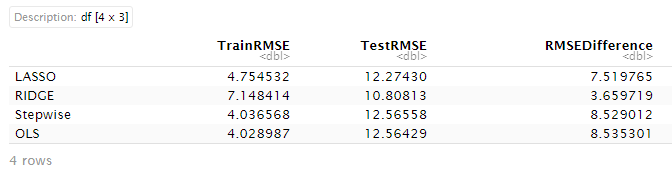
1. ***METHODS.***
2. **Model:**

The models that are using to predict are:

* Ordinary Least Squares or Linear Regression
* Stepwise Regression (Both Direction)
* LASSO Regression
* Ridge Regression
* Elastic Net Fit Regression
* Random Forest
* Gradient Boosting
* XG Boost

1. **Reasons:**

* Using Root Mean Square Error as metric to measure which model is the best the use.
* Firstly, we build OLS Regression – Old Least Square Regression which shows significantly high multi collinearity for 2 variables: Infant Death (VIF=173.31) and under-five deaths (VIF=171.65). After removing one of 2 variables about we got a model without Multi collinearity but still have overfitting problem: Train Root Mean Square Error (RMSE) is 4.03, and Test Root Mean Square Error is 12.57.
* We run Stepwise Regression– both direction: got over fitting problem: Train Root Mean Square Error is 4.04, and Test Root Mean Square Error is 12.57.
* Then, we decided to build Ridge Regression: the difference from Test RMSE and Train RMSE is reduced with Train Root Mean Square Error is 7.15, Test Root Mean Square Error is 10.8.
* Next, we build Lasso model, which yields RMSE even higher than that of Ridge model: Train Root Mean Square Error is 7.15, and Test Root Mean Square Error is 10.8.
* Consequently, we build Elastic Net Fit model but the Elastic fit model was computationally inefficient and could not process our nominal variable “Status”, hence we decided to reject the model move on with the Random Forest Model.
* The Random Forest model uses bagging techniques to build multiple decision trees to run the regression, here we get a Test Root Mean Square Error of 1.701
* Then we run the gradient Boosting Model, as it consecutively builds Decision trees in order to increase accuracy of the regression and is better than Random Forest, but we get a Test Root Mean Square Error of 1.79.
* Then we run the XGBoost model as well and we get a Test Root Mean Square Error of 1.87



1. ***ANALYSIS.***

## Slit the data into Train and Test sets

Because the formula for lasso and ridge regressions require x and y - X for all predictors variables and y for only response variables, I created x and y for both test and train sets.

## Ordinary Least Squares

ols<-lm(`Life expectancy`~., data=train)  
summary(ols)

##   
## Call:  
## lm(formula = `Life expectancy` ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.360 -2.280 -0.008 2.262 18.670   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.529e+01 7.916e-01 69.838 < 2e-16 \*\*\*  
## StatusDeveloping -1.600e+00 3.389e-01 -4.722 2.52e-06 \*\*\*  
## `Adult Mortality` -1.616e-02 1.008e-03 -16.035 < 2e-16 \*\*\*  
## `infant deaths` 8.895e-02 1.017e-02 8.744 < 2e-16 \*\*\*  
## Alcohol -6.673e-02 3.325e-02 -2.007 0.044903 \*   
## `percentage expenditure` 1.816e-04 9.360e-05 1.940 0.052484 .   
## Measles -1.400e-05 1.167e-05 -1.200 0.230382   
## BMI 3.917e-02 6.406e-03 6.114 1.20e-09 \*\*\*  
## `under-five deaths` -6.589e-02 7.479e-03 -8.810 < 2e-16 \*\*\*  
## Polio 2.014e-02 5.640e-03 3.572 0.000364 \*\*\*  
## `Total expenditure` 6.696e-02 4.379e-02 1.529 0.126408   
## Diphtheria 3.033e-02 5.653e-03 5.365 9.19e-08 \*\*\*  
## `HIV/AIDS` -4.821e-01 2.161e-02 -22.307 < 2e-16 \*\*\*  
## GDP 2.442e-05 1.342e-05 1.819 0.069043 .   
## Population -2.612e-10 2.381e-09 -0.110 0.912674   
## `thinness 1-19 years` -1.316e-01 5.988e-02 -2.197 0.028133 \*   
## `thinness 5-9 years` 1.219e-02 5.913e-02 0.206 0.836627   
## `Income composition of resources` 7.344e+00 7.952e-01 9.235 < 2e-16 \*\*\*  
## Schooling 7.366e-01 5.236e-02 14.066 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.964 on 1737 degrees of freedom  
## Multiple R-squared: 0.8329, Adjusted R-squared: 0.8311   
## F-statistic: 480.8 on 18 and 1737 DF, p-value: < 2.2e-16

coef(ols)

## (Intercept) StatusDeveloping   
## 5.528738e+01 -1.600082e+00   
## `Adult Mortality` `infant deaths`   
## -1.615569e-02 8.895109e-02   
## Alcohol `percentage expenditure`   
## -6.673469e-02 1.816363e-04   
## Measles BMI   
## -1.400029e-05 3.916635e-02   
## `under-five deaths` Polio   
## -6.589089e-02 2.014414e-02   
## `Total expenditure` Diphtheria   
## 6.695769e-02 3.033058e-02   
## `HIV/AIDS` GDP   
## -4.820919e-01 2.441751e-05   
## Population `thinness 1-19 years`   
## -2.611999e-10 -1.315645e-01   
## `thinness 5-9 years` `Income composition of resources`   
## 1.219401e-02 7.344121e+00   
## Schooling   
## 7.365510e-01

##Detect Multicolinearity  
vif(ols)

## Status `Adult Mortality`   
## 1.953618 1.832337   
## `infant deaths` Alcohol   
## 173.320468 1.914628   
## `percentage expenditure` Measles   
## 4.537044 1.395549   
## BMI `under-five deaths`   
## 1.812115 171.648700   
## Polio `Total expenditure`   
## 1.939396 1.174850   
## Diphtheria `HIV/AIDS`   
## 1.974800 1.494144   
## GDP Population   
## 4.468142 1.375003   
## `thinness 1-19 years` `thinness 5-9 years`   
## 7.778978 7.977238   
## `Income composition of resources` Schooling   
## 3.175007 3.624407

##Detect Outliers  
outlierTest(ols)

## rstudent unadjusted p-value Bonferroni p  
## 2030 -5.460977 5.4198e-08 9.5172e-05  
## 76 4.792724 1.7855e-06 3.1354e-03  
## 979 -4.684335 3.0273e-06 5.3159e-03  
## 77 4.617266 4.1741e-06 7.3296e-03  
## 75 4.331052 1.5688e-05 2.7548e-02  
## 2026 -4.313158 1.6999e-05 2.9850e-02  
## 304 4.273988 2.0243e-05 3.5546e-02  
## 79 4.250899 2.2423e-05 3.9375e-02  
## 80 4.199961 2.8050e-05 4.9256e-02

Infant Death and under-five deaths are those two have significantly high multicollinearity of 173.32 and 171.65 respectively. Thinness 1-19 years, thinness 5-9 years are those in critical level of multicollinearity with 7.78 and 7.98 respectively. Having those multicolinearity will lead to over fitting problems. I will remove the ‘infant death’ variable

##OLS without infant death variable  
ols1<-lm(`Life expectancy` ~ `Adult Mortality`+ `Status`+ `under-five deaths`+`HIV/AIDS`+ `Alcohol`+`thinness 5-9 years`+ `thinness 1-19 years`+ `percentage expenditure`+ BMI+ Polio+ `Total expenditure`+ Diphtheria+ GDP + `Income composition of resources`+ Schooling+Population,data=train)  
coef(ols1)

## (Intercept) `Adult Mortality`   
## 5.394166e+01 -1.664525e-02   
## StatusDeveloping `under-five deaths`   
## -1.528804e+00 -1.113379e-03   
## `HIV/AIDS` Alcohol   
## -4.899559e-01 -1.005594e-01   
## `thinness 5-9 years` `thinness 1-19 years`   
## 6.303816e-02 -1.605488e-01   
## `percentage expenditure` BMI   
## 2.011922e-04 4.075034e-02   
## Polio `Total expenditure`   
## 2.160898e-02 6.930420e-02   
## Diphtheria GDP   
## 3.582519e-02 1.915130e-05   
## `Income composition of resources` Schooling   
## 8.002620e+00 7.617560e-01   
## Population   
## 2.404177e-09

vif(ols1)

## `Adult Mortality` Status   
## 1.818640 1.952475   
## `under-five deaths` `HIV/AIDS`   
## 1.736793 1.489834   
## Alcohol `thinness 5-9 years`   
## 1.888825 7.899055   
## `thinness 1-19 years` `percentage expenditure`   
## 7.753964 4.534029   
## BMI Polio   
## 1.803090 1.937676   
## `Total expenditure` Diphtheria   
## 1.173850 1.949927   
## GDP `Income composition of resources`   
## 4.458219 3.145906   
## Schooling Population   
## 3.612785 1.352194

There are not any high collinearity variables. Only thinness 5-9 years and thinness 1-19 but these are still acceptable so We will keep them like this.

### Stepwise for OLS

bestmodel\_ols<-step(ols1, direction = 'both')

## Start: AIC=4928.02  
## `Life expectancy` ~ `Adult Mortality` + Status + `under-five deaths` +   
## `HIV/AIDS` + Alcohol + `thinness 5-9 years` + `thinness 1-19 years` +   
## `percentage expenditure` + BMI + Polio + `Total expenditure` +   
## Diphtheria + GDP + `Income composition of resources` + Schooling +   
## Population  
##   
## Df Sum of Sq RSS AIC  
## - Population 1 16.3 28521 4927.0  
## - `thinness 5-9 years` 1 18.0 28523 4927.1  
## - GDP 1 32.1 28537 4928.0  
## <none> 28505 4928.0  
## - `under-five deaths` 1 34.4 28539 4928.1  
## - `Total expenditure` 1 39.4 28544 4928.5  
## - `percentage expenditure` 1 72.7 28577 4930.5  
## - `thinness 1-19 years` 1 113.3 28618 4933.0  
## - Alcohol 1 145.7 28650 4935.0  
## - Polio 1 230.9 28736 4940.2  
## - Status 1 320.0 28825 4945.6  
## - BMI 1 639.1 29144 4965.0  
## - Diphtheria 1 639.1 29144 4965.0  
## - `Income composition of resources` 1 1606.1 30111 5022.3  
## - Schooling 1 3336.7 31841 5120.4  
## - `Adult Mortality` 1 4321.4 32826 5173.9  
## - `HIV/AIDS` 1 8100.6 36605 5365.2  
##   
## Step: AIC=4927.03  
## `Life expectancy` ~ `Adult Mortality` + Status + `under-five deaths` +   
## `HIV/AIDS` + Alcohol + `thinness 5-9 years` + `thinness 1-19 years` +   
## `percentage expenditure` + BMI + Polio + `Total expenditure` +   
## Diphtheria + GDP + `Income composition of resources` + Schooling  
##   
## Df Sum of Sq RSS AIC  
## - `thinness 5-9 years` 1 16.5 28537 4926.0  
## - `under-five deaths` 1 20.4 28541 4926.3  
## - GDP 1 31.3 28552 4927.0  
## <none> 28521 4927.0  
## - `Total expenditure` 1 38.3 28559 4927.4  
## + Population 1 16.3 28505 4928.0  
## - `percentage expenditure` 1 73.7 28595 4929.6  
## - `thinness 1-19 years` 1 109.7 28631 4931.8  
## - Alcohol 1 145.4 28666 4934.0  
## - Polio 1 230.5 28752 4939.2  
## - Status 1 315.8 28837 4944.4  
## - BMI 1 642.4 29163 4964.1  
## - Diphtheria 1 645.4 29166 4964.3  
## - `Income composition of resources` 1 1617.9 30139 5021.9  
## - Schooling 1 3349.1 31870 5120.0  
## - `Adult Mortality` 1 4323.8 32845 5172.9  
## - `HIV/AIDS` 1 8108.2 36629 5364.4  
##   
## Step: AIC=4926.04  
## `Life expectancy` ~ `Adult Mortality` + Status + `under-five deaths` +   
## `HIV/AIDS` + Alcohol + `thinness 1-19 years` + `percentage expenditure` +   
## BMI + Polio + `Total expenditure` + Diphtheria + GDP + `Income composition of resources` +   
## Schooling  
##   
## Df Sum of Sq RSS AIC  
## - `under-five deaths` 1 15.8 28553 4925.0  
## - GDP 1 30.3 28568 4925.9  
## <none> 28537 4926.0  
## - `Total expenditure` 1 35.2 28573 4926.2  
## + `thinness 5-9 years` 1 16.5 28521 4927.0  
## + Population 1 14.7 28523 4927.1  
## - `percentage expenditure` 1 74.2 28612 4928.6  
## - Alcohol 1 148.0 28685 4933.1  
## - `thinness 1-19 years` 1 192.8 28730 4935.9  
## - Polio 1 228.6 28766 4938.1  
## - Status 1 312.3 28850 4943.2  
## - BMI 1 626.6 29164 4962.2  
## - Diphtheria 1 651.2 29189 4963.7  
## - `Income composition of resources` 1 1625.5 30163 5021.3  
## - Schooling 1 3358.6 31896 5119.4  
## - `Adult Mortality` 1 4311.3 32849 5171.1  
## - `HIV/AIDS` 1 8097.9 36635 5362.7  
##   
## Step: AIC=4925.02  
## `Life expectancy` ~ `Adult Mortality` + Status + `HIV/AIDS` +   
## Alcohol + `thinness 1-19 years` + `percentage expenditure` +   
## BMI + Polio + `Total expenditure` + Diphtheria + GDP + `Income composition of resources` +   
## Schooling  
##   
## Df Sum of Sq RSS AIC  
## - GDP 1 30.9 28584 4924.9  
## <none> 28553 4925.0  
## - `Total expenditure` 1 36.1 28589 4925.2  
## + `under-five deaths` 1 15.8 28537 4926.0  
## + `thinness 5-9 years` 1 11.9 28541 4926.3  
## + Population 1 2.5 28551 4926.9  
## - `percentage expenditure` 1 73.2 28627 4927.5  
## - Alcohol 1 160.7 28714 4932.9  
## - Polio 1 237.4 28791 4937.6  
## - `thinness 1-19 years` 1 298.6 28852 4941.3  
## - Status 1 310.1 28863 4942.0  
## - BMI 1 625.5 29179 4961.1  
## - Diphtheria 1 659.1 29212 4963.1  
## - `Income composition of resources` 1 1613.3 30167 5019.5  
## - Schooling 1 3405.3 31959 5120.9  
## - `Adult Mortality` 1 4297.3 32851 5169.2  
## - `HIV/AIDS` 1 8083.9 36637 5360.8  
##   
## Step: AIC=4924.91  
## `Life expectancy` ~ `Adult Mortality` + Status + `HIV/AIDS` +   
## Alcohol + `thinness 1-19 years` + `percentage expenditure` +   
## BMI + Polio + `Total expenditure` + Diphtheria + `Income composition of resources` +   
## Schooling  
##   
## Df Sum of Sq RSS AIC  
## - `Total expenditure` 1 27.9 28612 4924.6  
## <none> 28584 4924.9  
## + GDP 1 30.9 28553 4925.0  
## + `under-five deaths` 1 16.4 28568 4925.9  
## + `thinness 5-9 years` 1 11.0 28573 4926.2  
## + Population 1 2.2 28582 4926.8  
## - Alcohol 1 176.5 28761 4933.7  
## - Polio 1 245.8 28830 4938.0  
## - `thinness 1-19 years` 1 300.0 28884 4941.3  
## - Status 1 316.5 28901 4942.2  
## - `percentage expenditure` 1 544.1 29128 4956.0  
## - BMI 1 647.3 29231 4962.2  
## - Diphtheria 1 662.1 29246 4963.1  
## - `Income composition of resources` 1 1670.0 30254 5022.6  
## - Schooling 1 3430.0 32014 5121.9  
## - `Adult Mortality` 1 4319.2 32903 5170.0  
## - `HIV/AIDS` 1 8065.2 36649 5359.4  
##   
## Step: AIC=4924.63  
## `Life expectancy` ~ `Adult Mortality` + Status + `HIV/AIDS` +   
## Alcohol + `thinness 1-19 years` + `percentage expenditure` +   
## BMI + Polio + Diphtheria + `Income composition of resources` +   
## Schooling  
##   
## Df Sum of Sq RSS AIC  
## <none> 28612 4924.6  
## + `Total expenditure` 1 27.9 28584 4924.9  
## + GDP 1 22.6 28589 4925.2  
## + `under-five deaths` 1 17.1 28595 4925.6  
## + `thinness 5-9 years` 1 8.9 28603 4926.1  
## + Population 1 1.8 28610 4926.5  
## - Alcohol 1 164.6 28777 4932.7  
## - Polio 1 243.8 28856 4937.5  
## - `thinness 1-19 years` 1 319.6 28932 4942.1  
## - Status 1 331.6 28944 4942.9  
## - `percentage expenditure` 1 568.8 29181 4957.2  
## - BMI 1 663.9 29276 4962.9  
## - Diphtheria 1 668.9 29281 4963.2  
## - `Income composition of resources` 1 1649.2 30261 5021.0  
## - Schooling 1 3472.0 32084 5123.7  
## - `Adult Mortality` 1 4319.7 32932 5169.5  
## - `HIV/AIDS` 1 8040.5 36653 5357.5

coef(bestmodel\_ols)

## (Intercept) `Adult Mortality`   
## 54.2357835261 -0.0166144174   
## StatusDeveloping `HIV/AIDS`   
## -1.5499515089 -0.4855396788   
## Alcohol `thinness 1-19 years`   
## -0.1051759331 -0.1213101716   
## `percentage expenditure` BMI   
## 0.0003161269 0.0410320606   
## Polio Diphtheria   
## 0.0221287645 0.0365905044   
## `Income composition of resources` Schooling   
## 8.0402964176 0.7734816689

summary(bestmodel\_ols)

##   
## Call:  
## lm(formula = `Life expectancy` ~ `Adult Mortality` + Status +   
## `HIV/AIDS` + Alcohol + `thinness 1-19 years` + `percentage expenditure` +   
## BMI + Polio + Diphtheria + `Income composition of resources` +   
## Schooling, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.9231 -2.2486 0.0452 2.3230 19.5286   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.424e+01 7.478e-01 72.530 < 2e-16 \*\*\*  
## `Adult Mortality` -1.661e-02 1.024e-03 -16.227 < 2e-16 \*\*\*  
## StatusDeveloping -1.550e+00 3.448e-01 -4.496 7.39e-06 \*\*\*  
## `HIV/AIDS` -4.855e-01 2.193e-02 -22.138 < 2e-16 \*\*\*  
## Alcohol -1.052e-01 3.321e-02 -3.167 0.00157 \*\*   
## `thinness 1-19 years` -1.213e-01 2.749e-02 -4.413 1.08e-05 \*\*\*  
## `percentage expenditure` 3.161e-04 5.369e-05 5.888 4.67e-09 \*\*\*  
## BMI 4.103e-02 6.450e-03 6.362 2.55e-10 \*\*\*  
## Polio 2.213e-02 5.740e-03 3.855 0.00012 \*\*\*  
## Diphtheria 3.659e-02 5.730e-03 6.385 2.19e-10 \*\*\*  
## `Income composition of resources` 8.040e+00 8.019e-01 10.026 < 2e-16 \*\*\*  
## Schooling 7.735e-01 5.317e-02 14.547 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.05 on 1744 degrees of freedom  
## Multiple R-squared: 0.8248, Adjusted R-squared: 0.8237   
## F-statistic: 746.4 on 11 and 1744 DF, p-value: < 2.2e-16

-Step wise function generates it best model with 11 variables that are StatusDeveloping, Adult Mortality, Alcohol, percentage expenditure, BMI, Polio, Diphtheria, HIV/AIDS, thinness 1-19 years, Income composition of resources, and Schooling. The average life expectancy is 55.2 years old.

* An increase in Adult Mortality, Alcohol, StatusDeveloping, HIV/AIDS, thinness 1-19 years will lead to a decrease in Life Expectancy an amount that equal to their corresponding coefficients. For example, an increase 1 people in Adult death causes a decrease in 0.0167 year of Life expectancy.
* percentage expenditure, BMI, Polio, Total expenditure, Diphtheria, GDP, Income composition of resources, and Schooling: have positive relationship with life expectancy

## Lasso Regression

#Alpha=1 for Ridge Regression

set.seed(123)  
cv.lasso<-cv.glmnet(train\_x, train\_y, nfolds = 10, alpha=1, family="gaussian")  
cv.lasso

##   
## Call: cv.glmnet(x = train\_x, y = train\_y, nfolds = 10, alpha = 1, family = "gaussian")   
##   
## Measure: Mean-Squared Error   
##   
## Lambda Index Measure SE Nonzero  
## min 0.0029 85 16.01 1.086 17  
## 1se 0.3664 33 17.05 1.236 11

###  
#lamda min- best lamda  
best\_lamda<-cv.lasso$lambda.min  
best\_lamda

## [1] 0.002903939

##   
#lamda 1se  
lamda\_1se<-cv.lasso$lambda.1se  
lamda\_1se

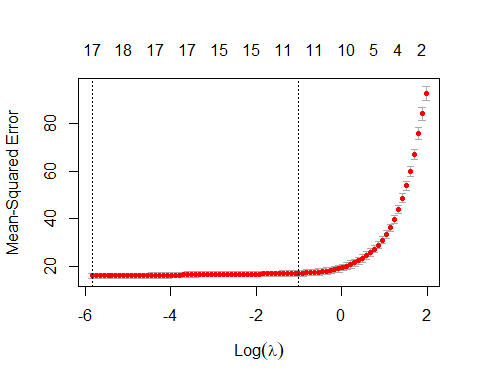
## [1] 0.3664356

* Cross-validation function will create 10 folds of the data, then take 9 for training and 1 for test. It will repeat the process until all the folds are already used as test and train.

-The best lamda is 0.0029 with 17 non-zero coefficients (17 out of 18 variables) in the model. The lamda.1se is 0.366 with 11 non-zero coefficients in the model.

#Plot of MSE by log lamda value

#Plot of MSE by log lamda value  
plot(cv.lasso)



## Values of log lamda  
log(cv.lasso$lambda.min)

## [1] -5.841687

log(cv.lasso$lambda.1se)

## [1] -1.003933

The dotted line on the left shows the optimal value of log lamda (close to -5.84) which minimizes the prediction error, and the number of variables (17) regardind to this lamda. The dotted line on the right represents the maximum the value of lamda (-1) within 1 standard error (1 se) of the minimum, and its number of variables (11) in the model.

##Fit Lasso model on lamda min

bestmodel\_l<-glmnet(train\_x, train\_y, alpha=1, lamda=best\_lamda)  
coef(cv.lasso, best\_lamda)

## 19 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 5.510895e+01  
## StatusDeveloping -1.576782e+00  
## `Adult Mortality` -1.622950e-02  
## `infant deaths` 7.615337e-02  
## Alcohol -6.936414e-02  
## `percentage expenditure` 1.836919e-04  
## Measles -1.309335e-05  
## BMI 3.915963e-02  
## `under-five deaths` -5.648971e-02  
## Polio 2.027689e-02  
## `Total expenditure` 6.527395e-02  
## Diphtheria 3.111089e-02  
## `HIV/AIDS` -4.827554e-01  
## GDP 2.358872e-05  
## Population .   
## `thinness 1-19 years` -1.252939e-01  
## `thinness 5-9 years` 7.928528e-03  
## `Income composition of resources` 7.442395e+00  
## Schooling 7.399359e-01

INTERPRET THE MODEL:

There are 17 predictors in this model, all of them except population, which is considered to have no impact on Life Expectancy. The average Life Expectancy is 55 years old. An increase in 1 unit of infant death, percentage expenditure, BMI, Polio, Total Expenditure, Diphtheria, thinness 5-9 years, Income composition of resources, and School would lead to a raise in Life Expectancy an amount that equal to their coefficients. For example, an increase in 1 percent of Income Composition of resources will raise 7.44 years in life expectancy.

In contrast, thinness 1-19 years, HIV/AIDS, under-five deaths, ALcohol, aldult, Mortality and Developing whose increases would cause a decrease in Life Expectancy an amount that equal to their coefficients. For example, if a country is Developing, it will decrease 1.58 year of life expectancy

I was so surprised that Infant deaths, Polio, and Diphtheria have positive impacts on Life Expectancy.

## Ridge Regression

#Alpha=0 for Ridge Regression

#Find the best lamda using cross-validation   
set.seed(123)  
cv.ridge<-cv.glmnet(train\_x, train\_y, nfolds = 10, alpha=0, family="gaussian")  
cv.ridge

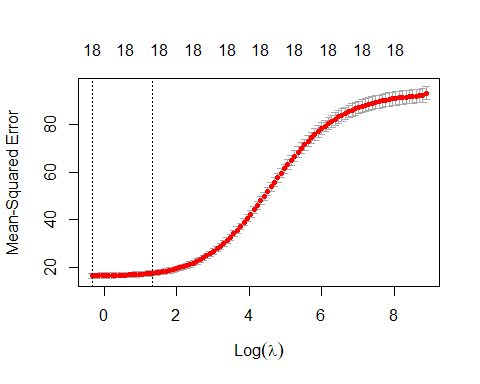
##   
## Call: cv.glmnet(x = train\_x, y = train\_y, nfolds = 10, alpha = 0, family = "gaussian")   
##   
## Measure: Mean-Squared Error   
##   
## Lambda Index Measure SE Nonzero  
## min 0.719 100 16.64 1.110 18  
## 1se 3.839 82 17.74 1.083 18

#lamda min- best lamda  
best\_lamda\_r<-cv.ridge$lambda.min  
#lamda 1se  
lamda\_1se\_r<-cv.ridge$lambda.1se

The best lamda, which is the optimal value to reduce the prectition error, is 0.719. The lamda.1se, which is the maximum lamda with 1 standard error of the best lamda, is 3.84. Both have 18 variables in their models.

#Plot of MSE by log lamda value

plot(cv.ridge)



log(best\_lamda\_r)

## [1] -0.329438

log(lamda\_1se\_r)

## [1] 1.345169

* The dotted line on the left shows the optimal value of log lamda (close to -0.33) which minimizes the prediction error. The dotted line on the right represents the maximum the value of lamda (1.35) within 1 standard error (1 se) of the minimum.
* Ridge Regression does not reduce the coefficients to zero so both models have 18 variables. ##Fit the ridge model for lamda min

bestmodel\_r<-glmnet(train\_x, train\_y, alpha=0, lamda=cv.ridge$lambda.min)  
coef(cv.ridge, best\_lamda\_r)

## 19 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 5.458182e+01  
## StatusDeveloping -1.527244e+00  
## `Adult Mortality` -1.637622e-02  
## `infant deaths` 3.240734e-03  
## Alcohol -6.584583e-02  
## `percentage expenditure` 1.821192e-04  
## Measles -1.049429e-05  
## BMI 4.213262e-02  
## `under-five deaths` -2.863445e-03  
## Polio 2.372175e-02  
## `Total expenditure` 6.226895e-02  
## Diphtheria 3.547745e-02  
## `HIV/AIDS` -4.611672e-01  
## GDP 2.434917e-05  
## Population 1.948735e-09  
## `thinness 1-19 years` -1.118103e-01  
## `thinness 5-9 years` -4.203134e-03  
## `Income composition of resources` 8.118195e+00  
## Schooling 6.741227e-01

INTERPRET THE MODEL:

* There are all 18 variables in this model. The average life expectancy is 56.6 years old. Again, an increase in infant death, BMI, Polio, percentage expenditure, Total Expenditure, Diphtheria, population, Income composition of resources, and School will lead to an increase in life expectancy an amount equal to its coefficients. For example, an increase in 1 people of Polio will raise 0.024 years in life expectancy.
* Different from Lasso model, thinness 5-9 years variable in this model has negative impact on life expectancy.
* The rest of the variables have negative relationship to Life Expectancy.

## Determine the performance of OLS, Ridge, and Lasso models on train and test sets.

### Lasso

##Running on train set  
preds.train.l<-predict(bestmodel\_l, newx=train\_x)  
train.rmse.l <- rmse(train\_y,preds.train.l)  
##Running on test sets  
preds.test.l<-predict(bestmodel\_l, newx = test\_x)  
test.rmse.l <- rmse(train\_y,preds.test.l)

## Warning in actual - predicted: longer object length is not a multiple of shorter  
## object length

##Compare the RMSE  
train.rmse.l

## [1] 4.754532

test.rmse.l

## [1] 12.2743

* In this Lasso model, the Train Root Mean Square Error is 4.8, while the Test Root Mean Square Error is 12.3

### Ridge

##Running on train set  
preds.train.r<-predict(bestmodel\_r, newx=train\_x)  
train.rmse.r <- rmse(train\_y,preds.train.r)  
##Running on test sets  
preds.test.r<-predict(bestmodel\_r, newx= test\_x)  
test.rmse.r <- rmse(train\_y,preds.test.r)

## Warning in actual - predicted: longer object length is not a multiple of shorter  
## object length

##Compare the RMSE  
train.rmse.r

## [1] 7.148414

test.rmse.r

## [1] 10.80813

* In this Ridge model, the Train Root Mean Square Error is 7.15, while the Test Root Mean Square Error is 10.8.

### OLS

preds.train<-predict(ols1, new=train)  
train.rmse <- rmse(train\_y,preds.train)  
##Running on test sets  
preds.test<-predict(ols1, new= test)  
test.rmse <- rmse(train\_y,preds.test)

## Warning in actual - predicted: longer object length is not a multiple of shorter  
## object length

##Compare the RMSE  
train.rmse

## [1] 4.028987

test.rmse

## [1] 12.56429

* In this Old Least Square model, the Train Root Mean Square Error is 4.03, while the Test Root Mean Square Error is 12.57

### OLS Stepwise

preds.train\_st<-predict(bestmodel\_ols, new=train)  
train.rmse\_st <- rmse(train\_y,preds.train\_st)  
##Running on test sets  
preds.test\_st<-predict(bestmodel\_ols, new= test)  
test.rmse\_st <- rmse(train\_y,preds.test\_st)

## Warning in actual - predicted: longer object length is not a multiple of shorter  
## object length

##Compare the RMSE  
train.rmse\_st

## [1] 4.036568

test.rmse\_st

## [1] 12.56558

* In this Old Least Square model, the Train Root Mean Square Error is 4.04, while the Test Root Mean Square Error is 12.57.

TrainRMSE<-c(train.rmse.l, train.rmse.r, train.rmse\_st, train.rmse)  
TestRMSE<-c(test.rmse.l, test.rmse.r, test.rmse\_st, test.rmse)  
RMSEDifference<-TestRMSE-TrainRMSE  
  
compare<-data.frame(TrainRMSE, TestRMSE, RMSEDifference)  
 rownames(compare)<-c("LASSO","RIDGE","Stepwise","OLS")  
 compare

## TrainRMSE TestRMSE RMSEDifference  
## LASSO 4.754532 12.27430 7.519765  
## RIDGE 7.148414 10.80813 3.659719  
## Stepwise 4.036568 12.56558 8.529012  
## OLS 4.028987 12.56429 8.535301

# As we can observe we did not get any satisfactory results in the simple models like Linear Regression, Lasso, Ridge and Stepwise. We decide to try some more complex and Black-box Models like Elastic net fit, Random Forest which is a type of Bagging Model and Boosting Models like Gradient Boosting and XGBoost.

Elastic Net Fit:

The elastic net model combines the L1 and L2 penalty terms:

1

Here we have a parameter alpha that blends the two penalty terms together.  When alpha equals 0 we get Ridge regression.  If alpha is set to 1 then we have the LASSO model.  The lambda parameter is the shrinkage coefficient.

# Model Building : Elastic Net Regression

control <- trainControl(method = “repeatedcv”, number = 5, repeats = 5, search = “random”, verboseIter = TRUE)

# Training ELastic Net Regression model

elastic\_model <- train(Life\_exp ~ ., data = as.matrix(dfLE1), method = “glmnet”, preProcess = c(“center”, “scale”), tuneLength = 25, trControl = control)

elastic\_model

# Model Prediction

x\_hat\_pre <- predict(elastic\_model, Y) x\_hat\_pre

# Multiple R-squared

rsq <- cor(X, x\_hat\_pre)^2 rsq

# Plot

plot(elastic\_model, main = “Elastic Net Regression”) Output:

Result: The Elastic fit model was computationally inefficient and could not process our nominal variable “Status”, hence we decided to reject the model move on with the Random Forest Model.

####Feature Selection

#Checking Variance  
var(dfLE1)

## Warning in var(dfLE1): NAs introduced by coercion

## Status Life expectancy Adult Mortality  
## Status NA NA NA  
## Life expectancy NA 9.227790e+01 -8.314218e+02  
## Adult Mortality NA -8.314218e+02 1.583780e+04  
## infant deaths NA -2.228491e+02 1.053617e+03  
## Alcohol NA 1.398334e+01 -8.413500e+01  
## percentage expenditure NA 8.322527e+03 -6.772151e+04  
## Measles NA -1.559895e+04 1.766158e+04  
## BMI NA 1.087351e+02 -9.447699e+02  
## under-five deaths NA -3.450295e+02 1.752461e+03  
## Polio NA 1.018480e+02 -7.538939e+02  
## Total expenditure NA 3.943684e+00 -2.224162e+01  
## Diphtheria NA 1.049853e+02 -7.383590e+02  
## HIV/AIDS NA -2.951550e+01 3.655827e+02  
## GDP NA 6.304021e+04 -5.290125e+05  
## Population NA -1.703633e+07 -4.364778e+07  
## thinness 1-19 years NA -2.044982e+01 1.649712e+02  
## thinness 5-9 years NA -2.069158e+01 1.731902e+02  
## Income composition of resources NA 1.495183e+00 -1.231395e+01  
## Schooling NA 2.439071e+01 -1.909882e+02  
## infant deaths Alcohol  
## Status NA NA  
## Life expectancy -2.228491e+02 1.398334e+01  
## Adult Mortality 1.053617e+03 -8.413500e+01  
## infant deaths 1.589603e+04 -5.005152e+01  
## Alcohol -5.005152e+01 1.517793e+01  
## percentage expenditure -2.420206e+04 3.143634e+03  
## Measles 7.261383e+05 -1.283641e+03  
## BMI -5.716397e+02 2.496914e+01  
## under-five deaths 2.153946e+04 -6.553724e+01  
## Polio -4.886029e+02 1.897625e+01  
## Total expenditure -3.656883e+01 2.677870e+00  
## Diphtheria -4.944892e+02 1.950539e+01  
## HIV/AIDS 1.102193e+01 -8.001041e-01  
## GDP -1.967857e+05 1.886248e+04  
## Population 4.099618e+09 -7.233462e+06  
## thinness 1-19 years 2.716555e+02 -7.084671e+00  
## thinness 5-9 years 2.809385e+02 -7.009669e+00  
## Income composition of resources -4.090293e+00 3.676208e-01  
## Schooling -8.675181e+01 7.081459e+00  
## percentage expenditure Measles  
## Status NA NA  
## Life expectancy 8.322527e+03 -1.559895e+04  
## Adult Mortality -6.772151e+04 1.766158e+04  
## infant deaths -2.420206e+04 7.261383e+05  
## Alcohol 3.143634e+03 -1.283641e+03  
## percentage expenditure 4.518606e+06 -1.443119e+06  
## Measles -1.443119e+06 1.220297e+08  
## BMI 1.046037e+04 -3.786853e+04  
## under-five deaths -3.383580e+04 1.001081e+06  
## Polio 8.145734e+03 -3.110053e+04  
## Total expenditure 1.029135e+03 -2.323828e+03  
## Diphtheria 7.881200e+03 -3.182168e+04  
## HIV/AIDS -1.269416e+03 1.729900e+03  
## GDP 2.710208e+07 -1.218041e+07  
## Population -3.376235e+09 1.723934e+11  
## thinness 1-19 years -2.568703e+03 1.163836e+04  
## thinness 5-9 years -2.645220e+03 1.167789e+04  
## Income composition of resources 1.780965e+02 -3.239638e+02  
## Schooling 2.876610e+03 -5.463560e+03  
## BMI under-five deaths Polio  
## Status NA NA NA  
## Life expectancy 1.087351e+02 -3.450295e+02 1.018480e+02  
## Adult Mortality -9.447699e+02 1.752461e+03 -7.538939e+02  
## infant deaths -5.716397e+02 2.153946e+04 -4.886029e+02  
## Alcohol 2.496914e+01 -6.553724e+01 1.897625e+01  
## percentage expenditure 1.046037e+04 -3.383580e+04 8.145734e+03  
## Measles -3.786853e+04 1.001081e+06 -3.110053e+04  
## BMI 3.952293e+02 -8.108789e+02 1.286910e+02  
## under-five deaths -8.108789e+02 2.938452e+04 -7.353303e+02  
## Polio 1.286910e+02 -7.353303e+02 5.460228e+02  
## Total expenditure 8.784836e+00 -4.989412e+01 6.312756e+00  
## Diphtheria 1.276120e+02 -7.547570e+02 3.724804e+02  
## HIV/AIDS -2.633929e+01 2.604983e+01 -1.987680e+01  
## GDP 8.472230e+04 -2.764719e+05 7.153546e+04  
## Population -8.514723e+07 5.450486e+09 -5.990664e+07  
## thinness 1-19 years -4.690969e+01 3.705438e+02 -2.235401e+01  
## thinness 5-9 years -4.866436e+01 3.822979e+02 -2.318946e+01  
## Income composition of resources 2.191081e+00 -6.236418e+00 1.907853e+00  
## Schooling 3.675179e+01 -1.272198e+02 3.318214e+01  
## Total expenditure Diphtheria HIV/AIDS  
## Status NA NA NA  
## Life expectancy 3.943684e+00 1.049853e+02 -2.951550e+01  
## Adult Mortality -2.224162e+01 -7.383590e+02 3.655827e+02  
## infant deaths -3.656883e+01 -4.944892e+02 1.102193e+01  
## Alcohol 2.677870e+00 1.950539e+01 -8.001041e-01  
## percentage expenditure 1.029135e+03 7.881200e+03 -1.269416e+03  
## Measles -2.323828e+03 -3.182168e+04 1.729900e+03  
## BMI 8.784836e+00 1.276120e+02 -2.633929e+01  
## under-five deaths -4.989412e+01 -7.547570e+02 2.604983e+01  
## Polio 6.312756e+00 3.724804e+02 -1.987680e+01  
## Total expenditure 5.452978e+00 6.846711e+00 3.200876e-01  
## Diphtheria 6.846711e+00 5.532528e+02 -2.085547e+01  
## HIV/AIDS 3.200876e-01 -2.085547e+01 2.934929e+01  
## GDP 4.404366e+03 6.780950e+04 -1.062939e+04  
## Population -9.110021e+06 -4.589651e+07 -7.410878e+06  
## thinness 1-19 years -2.560153e+00 -2.341082e+01 4.913344e+00  
## thinness 5-9 years -2.709240e+00 -2.352617e+01 5.111705e+00  
## Income composition of resources 8.057832e-02 2.028767e+00 -2.942737e-01  
## Schooling 1.870580e+00 3.388374e+01 -4.133641e+00  
## GDP Population  
## Status NA NA  
## Life expectancy 6.304021e+04 -1.703633e+07  
## Adult Mortality -5.290125e+05 -4.364778e+07  
## infant deaths -1.967857e+05 4.099618e+09  
## Alcohol 1.886248e+04 -7.233462e+06  
## percentage expenditure 2.710208e+07 -3.376235e+09  
## Measles -1.218041e+07 1.723934e+11  
## BMI 8.472230e+04 -8.514723e+07  
## under-five deaths -2.764719e+05 5.450486e+09  
## Polio 7.153546e+04 -5.990664e+07  
## Total expenditure 4.404366e+03 -9.110021e+06  
## Diphtheria 6.780950e+04 -4.589651e+07  
## HIV/AIDS -1.062939e+04 -7.410878e+06  
## GDP 2.145865e+08 -2.958492e+10  
## Population -2.958492e+10 3.400580e+15  
## thinness 1-19 years -1.811021e+04 6.604386e+07  
## thinness 5-9 years -1.879078e+04 6.689846e+07  
## Income composition of resources 1.388269e+03 -1.735477e+05  
## Schooling 2.166886e+04 -7.550680e+06  
## thinness 1-19 years thinness 5-9 years  
## Status NA NA  
## Life expectancy -2.044982e+01 -2.069158e+01  
## Adult Mortality 1.649712e+02 1.731902e+02  
## infant deaths 2.716555e+02 2.809385e+02  
## Alcohol -7.084671e+00 -7.009669e+00  
## percentage expenditure -2.568703e+03 -2.645220e+03  
## Measles 1.163836e+04 1.167789e+04  
## BMI -4.690969e+01 -4.866436e+01  
## under-five deaths 3.705438e+02 3.822979e+02  
## Polio -2.235401e+01 -2.318946e+01  
## Total expenditure -2.560153e+00 -2.709240e+00  
## Diphtheria -2.341082e+01 -2.352617e+01  
## HIV/AIDS 4.913344e+00 5.111705e+00  
## GDP -1.811021e+04 -1.879078e+04  
## Population 6.604386e+07 6.689846e+07  
## thinness 1-19 years 1.991407e+01 1.895268e+01  
## thinness 5-9 years 1.895268e+01 2.073686e+01  
## Income composition of resources -4.086243e-01 -4.061399e-01  
## Schooling -7.018885e+00 -6.999899e+00  
## Income composition of resources Schooling  
## Status NA NA  
## Life expectancy 1.495183e+00 2.439071e+01  
## Adult Mortality -1.231395e+01 -1.909882e+02  
## infant deaths -4.090293e+00 -8.675181e+01  
## Alcohol 3.676208e-01 7.081459e+00  
## percentage expenditure 1.780965e+02 2.876610e+03  
## Measles -3.239638e+02 -5.463560e+03  
## BMI 2.191081e+00 3.675179e+01  
## under-five deaths -6.236418e+00 -1.272198e+02  
## Polio 1.907853e+00 3.318214e+01  
## Total expenditure 8.057832e-02 1.870580e+00  
## Diphtheria 2.028767e+00 3.388374e+01  
## HIV/AIDS -2.942737e-01 -4.133641e+00  
## GDP 1.388269e+03 2.166886e+04  
## Population -1.735477e+05 -7.550680e+06  
## thinness 1-19 years -4.086243e-01 -7.018885e+00  
## thinness 5-9 years -4.061399e-01 -6.999899e+00  
## Income composition of resources 4.498763e-02 5.750990e-01  
## Schooling 5.750990e-01 1.149020e+01

#None of the variables have Variance=0  
#Check for Multicollinearty  
library(mctest)  
X <- dfLE1 %>% select\_if(is.numeric)  
Y <- X$`Life expectancy`  
imcdiag (ols)

##   
## Call:  
## imcdiag(mod = ols)  
##   
##   
## All Individual Multicollinearity Diagnostics Result  
##   
## VIF TOL Wi Fi Leamer  
## StatusDeveloping 1.9536 0.5119 97.4934 103.6464 0.7155  
## `Adult Mortality` 1.8323 0.5458 85.0942 90.4646 0.7387  
## `infant deaths` 173.3205 0.0058 17617.2337 18729.0809 0.0760  
## Alcohol 1.9146 0.5223 93.5073 99.4087 0.7227  
## `percentage expenditure` 4.5370 0.2204 361.6108 384.4325 0.4695  
## Measles 1.3955 0.7166 40.4391 42.9912 0.8465  
## BMI 1.8121 0.5518 83.0268 88.2668 0.7429  
## `under-five deaths` 171.6487 0.0058 17446.3200 18547.3805 0.0763  
## Polio 1.9394 0.5156 96.0395 102.1006 0.7181  
## `Total expenditure` 1.1748 0.8512 17.8758 19.0040 0.9226  
## Diphtheria 1.9748 0.5064 99.6590 105.9486 0.7116  
## `HIV/AIDS` 1.4941 0.6693 50.5190 53.7073 0.8181  
## GDP 4.4681 0.2238 354.5666 376.9437 0.4731  
## Population 1.3750 0.7273 38.3385 40.7581 0.8528  
## `thinness 1-19 years` 7.7790 0.1286 693.0508 736.7902 0.3585  
## `thinness 5-9 years` 7.9772 0.1254 713.3200 758.3386 0.3541  
## `Income composition of resources` 3.1750 0.3150 222.3625 236.3961 0.5612  
## Schooling 3.6244 0.2759 268.3070 285.2402 0.5253  
## CVIF Klein IND1 IND2  
## StatusDeveloping -0.1081 0 0.0050 0.8304  
## `Adult Mortality` -0.1014 0 0.0053 0.7727  
## `infant deaths` -9.5892 1 0.0001 1.6913  
## Alcohol -0.1059 0 0.0051 0.8126  
## `percentage expenditure` -0.2510 0 0.0022 1.3262  
## Measles -0.0772 0 0.0070 0.4822  
## BMI -0.1003 0 0.0054 0.7624  
## `under-five deaths` -9.4967 1 0.0001 1.6912  
## Polio -0.1073 0 0.0050 0.8240  
## `Total expenditure` -0.0650 0 0.0083 0.2532  
## Diphtheria -0.1093 0 0.0050 0.8397  
## `HIV/AIDS` -0.0827 0 0.0065 0.5626  
## GDP -0.2472 0 0.0022 1.3204  
## Population -0.0761 0 0.0071 0.4639  
## `thinness 1-19 years` -0.4304 1 0.0013 1.4824  
## `thinness 5-9 years` -0.4414 1 0.0012 1.4879  
## `Income composition of resources` -0.1757 0 0.0031 1.1653  
## Schooling -0.2005 0 0.0027 1.2318  
##   
## 1 --> COLLINEARITY is detected by the test   
## 0 --> COLLINEARITY is not detected by the test  
##   
## `percentage expenditure` , Measles , `Total expenditure` , GDP , Population , `thinness 5-9 years` , coefficient(s) are non-significant may be due to multicollinearity  
##   
## R-square of y on all x: 0.8329   
##   
## \* use method argument to check which regressors may be the reason of collinearity  
## ===================================

#without Infant death variable  
ols\_1<-lm(`Life expectancy`~., data=train[,-4])  
summary(ols\_1)

##   
## Call:  
## lm(formula = `Life expectancy` ~ ., data = train[, -4])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.4905 -2.3131 0.0039 2.3493 19.4746   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.398e+01 7.942e-01 67.974 < 2e-16 \*\*\*  
## StatusDeveloping -1.529e+00 3.460e-01 -4.417 1.06e-05 \*\*\*  
## `Adult Mortality` -1.668e-02 1.027e-03 -16.240 < 2e-16 \*\*\*  
## Alcohol -1.003e-01 3.374e-02 -2.974 0.002980 \*\*   
## `percentage expenditure` 2.017e-04 9.559e-05 2.110 0.035025 \*   
## Measles -7.161e-06 1.189e-05 -0.602 0.547152   
## BMI 4.049e-02 6.542e-03 6.189 7.54e-10 \*\*\*  
## `under-five deaths` -8.973e-04 8.481e-04 -1.058 0.290207   
## Polio 2.159e-02 5.758e-03 3.750 0.000183 \*\*\*  
## `Total expenditure` 6.853e-02 4.473e-02 1.532 0.125681   
## Diphtheria 3.586e-02 5.739e-03 6.249 5.20e-10 \*\*\*  
## `HIV/AIDS` -4.895e-01 2.206e-02 -22.189 < 2e-16 \*\*\*  
## GDP 1.906e-05 1.370e-05 1.391 0.164285   
## Population 2.416e-09 2.412e-09 1.001 0.316789   
## `thinness 1-19 years` -1.609e-01 6.107e-02 -2.634 0.008501 \*\*   
## `thinness 5-9 years` 6.231e-02 6.011e-02 1.037 0.300097   
## `Income composition of resources` 7.992e+00 8.088e-01 9.881 < 2e-16 \*\*\*  
## Schooling 7.612e-01 5.341e-02 14.252 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.049 on 1738 degrees of freedom  
## Multiple R-squared: 0.8255, Adjusted R-squared: 0.8238   
## F-statistic: 483.6 on 17 and 1738 DF, p-value: < 2.2e-16

summary(ols)

##   
## Call:  
## lm(formula = `Life expectancy` ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.360 -2.280 -0.008 2.262 18.670   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.529e+01 7.916e-01 69.838 < 2e-16 \*\*\*  
## StatusDeveloping -1.600e+00 3.389e-01 -4.722 2.52e-06 \*\*\*  
## `Adult Mortality` -1.616e-02 1.008e-03 -16.035 < 2e-16 \*\*\*  
## `infant deaths` 8.895e-02 1.017e-02 8.744 < 2e-16 \*\*\*  
## Alcohol -6.673e-02 3.325e-02 -2.007 0.044903 \*   
## `percentage expenditure` 1.816e-04 9.360e-05 1.940 0.052484 .   
## Measles -1.400e-05 1.167e-05 -1.200 0.230382   
## BMI 3.917e-02 6.406e-03 6.114 1.20e-09 \*\*\*  
## `under-five deaths` -6.589e-02 7.479e-03 -8.810 < 2e-16 \*\*\*  
## Polio 2.014e-02 5.640e-03 3.572 0.000364 \*\*\*  
## `Total expenditure` 6.696e-02 4.379e-02 1.529 0.126408   
## Diphtheria 3.033e-02 5.653e-03 5.365 9.19e-08 \*\*\*  
## `HIV/AIDS` -4.821e-01 2.161e-02 -22.307 < 2e-16 \*\*\*  
## GDP 2.442e-05 1.342e-05 1.819 0.069043 .   
## Population -2.612e-10 2.381e-09 -0.110 0.912674   
## `thinness 1-19 years` -1.316e-01 5.988e-02 -2.197 0.028133 \*   
## `thinness 5-9 years` 1.219e-02 5.913e-02 0.206 0.836627   
## `Income composition of resources` 7.344e+00 7.952e-01 9.235 < 2e-16 \*\*\*  
## Schooling 7.366e-01 5.236e-02 14.066 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.964 on 1737 degrees of freedom  
## Multiple R-squared: 0.8329, Adjusted R-squared: 0.8311   
## F-statistic: 480.8 on 18 and 1737 DF, p-value: < 2.2e-16

imcdiag(ols\_1)

##   
## Call:  
## imcdiag(mod = ols\_1)  
##   
##   
## All Individual Multicollinearity Diagnostics Result  
##   
## VIF TOL Wi Fi Leamer  
## StatusDeveloping 1.9525 0.5122 103.5225 110.4875 0.7157  
## `Adult Mortality` 1.8258 0.5477 89.7488 95.7871 0.7401  
## Alcohol 1.8891 0.5294 96.6296 103.1308 0.7276  
## `percentage expenditure` 4.5343 0.2205 384.1374 409.9821 0.4696  
## Measles 1.3893 0.7198 42.3098 45.1564 0.8484  
## BMI 1.8111 0.5521 88.1571 94.0883 0.7431  
## `under-five deaths` 2.1154 0.4727 121.2308 129.3873 0.6875  
## Polio 1.9377 0.5161 101.9191 108.7762 0.7184  
## `Total expenditure` 1.1748 0.8512 19.0018 20.2803 0.9226  
## Diphtheria 1.9501 0.5128 103.2647 110.2124 0.7161  
## `HIV/AIDS` 1.4919 0.6703 53.4601 57.0569 0.8187  
## GDP 4.4588 0.2243 375.9301 401.2227 0.4736  
## Population 1.3523 0.7395 38.2883 40.8643 0.8599  
## `thinness 1-19 years` 7.7546 0.1290 734.1395 783.5325 0.3591  
## `thinness 5-9 years` 7.9023 0.1265 750.1920 800.6650 0.3557  
## `Income composition of resources` 3.1475 0.3177 233.4023 249.1056 0.5637  
## Schooling 3.6139 0.2767 284.0996 303.2138 0.5260  
## CVIF Klein IND1 IND2  
## StatusDeveloping -0.1140 0 0.0047 0.9132  
## `Adult Mortality` -0.1066 0 0.0050 0.8466  
## Alcohol -0.1103 0 0.0049 0.8810  
## `percentage expenditure` -0.2648 0 0.0020 1.4591  
## Measles -0.0811 0 0.0066 0.5245  
## BMI -0.1058 0 0.0051 0.8383  
## `under-five deaths` -0.1236 0 0.0043 0.9870  
## Polio -0.1132 0 0.0047 0.9059  
## `Total expenditure` -0.0686 0 0.0078 0.2786  
## Diphtheria -0.1139 0 0.0047 0.9120  
## `HIV/AIDS` -0.0871 0 0.0062 0.6172  
## GDP -0.2604 0 0.0021 1.4521  
## Population -0.0790 0 0.0068 0.4877  
## `thinness 1-19 years` -0.4529 1 0.0012 1.6305  
## `thinness 5-9 years` -0.4615 1 0.0012 1.6351  
## `Income composition of resources` -0.1838 0 0.0029 1.2772  
## Schooling -0.2111 0 0.0025 1.3540  
##   
## 1 --> COLLINEARITY is detected by the test   
## 0 --> COLLINEARITY is not detected by the test  
##   
## Measles , `under-five deaths` , `Total expenditure` , GDP , Population , `thinness 5-9 years` , coefficient(s) are non-significant may be due to multicollinearity  
##   
## R-square of y on all x: 0.8255   
##   
## \* use method argument to check which regressors may be the reason of collinearity  
## ===================================

Random Forest: Random forest is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

# Random Forest   
library(randomForest)

## randomForest 4.7-1

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

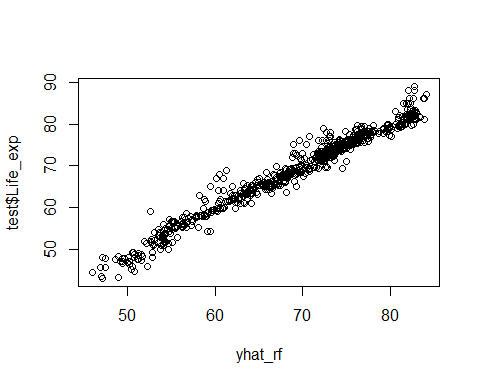
## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

set.seed(1)  
  
#Bagging  
names(train) <-c("Status","Life\_exp","Adl\_Mor","Inf\_Dth","Alcohol","Per\_exp","Measles","BMI","U5\_dth","Polio",  
 "Tot\_Exp","dip","HIV","GDP","POP","thin\_1\_19","thin\_5\_9","Icm\_comp\_res","schooling")  
  
names(test) <-c("Status","Life\_exp","Adl\_Mor","Inf\_Dth","Alcohol","Per\_exp","Measles","BMI","U5\_dth","Polio",  
 "Tot\_Exp","dip","HIV","GDP","POP","thin\_1\_19","thin\_5\_9","Icm\_comp\_res","schooling")  
  
  
rf\_model <-randomForest(`Life\_exp` ~. , data=train, mtry = 4, importance= TRUE)  
rf\_model

##   
## Call:  
## randomForest(formula = Life\_exp ~ ., data = train, mtry = 4, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 4  
##   
## Mean of squared residuals: 4.262069  
## % Var explained: 95.42

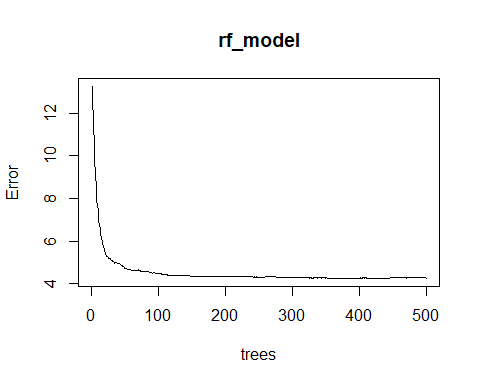
yhat\_rf = predict (rf\_model , newdata=test)  
plot(yhat\_rf , test$Life\_exp)



mean((yhat\_rf -test$Life\_exp)^2)

## [1] 2.999254

plot(rf\_model)



importance(rf\_model)

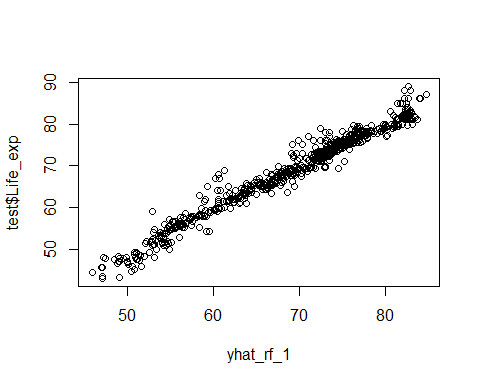
## %IncMSE IncNodePurity  
## Status 10.68947 1585.6677  
## Adl\_Mor 31.65092 23087.7682  
## Inf\_Dth 17.82266 5391.7547  
## Alcohol 29.24669 2054.5042  
## Per\_exp 16.20598 1887.8693  
## Measles 14.23870 952.0879  
## BMI 21.76904 10132.3172  
## U5\_dth 18.08368 7063.6858  
## Polio 17.84996 3249.5739  
## Tot\_Exp 24.19462 1510.9109  
## dip 18.27452 3332.5107  
## HIV 35.34915 33971.0921  
## GDP 19.14288 4160.0268  
## POP 15.39591 934.8750  
## thin\_1\_19 24.40906 4594.3035  
## thin\_5\_9 27.09006 4826.3928  
## Icm\_comp\_res 29.52846 36274.8354  
## schooling 22.66964 17264.1506

RESULT: Here we see that on test data we get a MSE value of 2.99, hence we try to improve this error term by tuning the hyperparameters used in Random forest model.

#Tuning hyperparameters  
  
rf\_1\_model <-randomForest(`Life\_exp` ~. , data=train, mtry = 4, importance= TRUE,ntree=600,nodesize=3)  
rf\_1\_model

##   
## Call:  
## randomForest(formula = Life\_exp ~ ., data = train, mtry = 4, importance = TRUE, ntree = 600, nodesize = 3)   
## Type of random forest: regression  
## Number of trees: 600  
## No. of variables tried at each split: 4  
##   
## Mean of squared residuals: 4.163936  
## % Var explained: 95.52

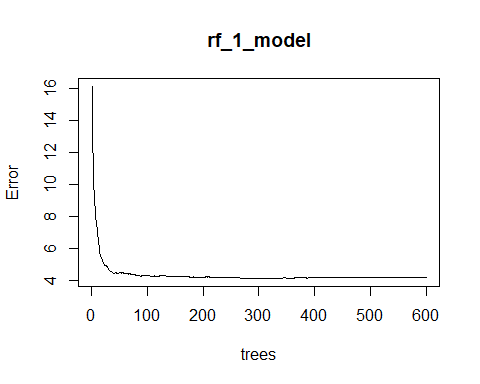
yhat\_rf\_1 = predict (rf\_1\_model , newdata=test)  
plot(yhat\_rf\_1 , test$Life\_exp)



mean((yhat\_rf\_1 -test$Life\_exp)^2)

## [1] 2.90288

plot(rf\_1\_model)



importance(rf\_1\_model)

## %IncMSE IncNodePurity  
## Status 11.12757 1623.4080  
## Adl\_Mor 34.94143 25106.7258  
## Inf\_Dth 18.42933 4855.3431  
## Alcohol 31.93746 2064.0451  
## Per\_exp 18.28801 1940.0110  
## Measles 10.71917 883.9178  
## BMI 22.26179 9972.2515  
## U5\_dth 14.35267 8014.3624  
## Polio 19.71144 3023.8158  
## Tot\_Exp 23.85517 1561.2494  
## dip 21.70106 3259.7057  
## HIV 36.47343 30114.3424  
## GDP 19.03469 4386.8255  
## POP 18.63410 956.3218  
## thin\_1\_19 24.69589 4567.4530  
## thin\_5\_9 29.35130 4652.1204  
## Icm\_comp\_res 34.61046 36239.3312  
## schooling 22.86814 19746.1335

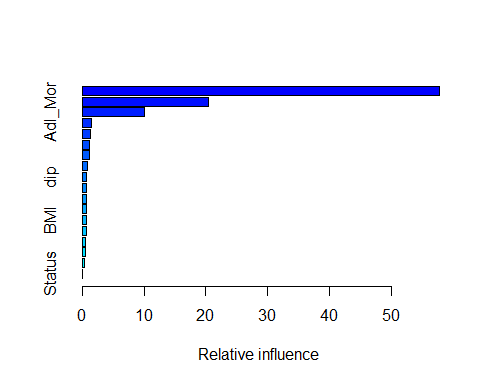
RESULT POST TUNING: Here we can see we were able to reduce the MSE to 2.90 after hypertuning. Hence, we take this as our final MSE for Random Forest Model.

*Gradient Boosting:* Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

library(gbm)

## Loaded gbm 2.1.8

set.seed(1)  
train$Status<-as.factor(train$Status)  
test$Status <-as.factor(test$Status)  
boost\_model=gbm(Life\_exp∼.,data=train, distribution="gaussian",n.trees=500, interaction.depth=4)  
summary(boost\_model)



## var rel.inf  
## Icm\_comp\_res Icm\_comp\_res 57.81615622  
## HIV HIV 20.45263458  
## Adl\_Mor Adl\_Mor 10.11678681  
## thin\_5\_9 thin\_5\_9 1.52623661  
## schooling schooling 1.37419493  
## U5\_dth U5\_dth 1.22018325  
## Tot\_Exp Tot\_Exp 1.11277899  
## Alcohol Alcohol 0.83711395  
## dip dip 0.75924322  
## Polio Polio 0.72279962  
## Inf\_Dth Inf\_Dth 0.69440167  
## thin\_1\_19 thin\_1\_19 0.65511217  
## BMI BMI 0.64242713  
## POP POP 0.63877808  
## Per\_exp Per\_exp 0.59279359  
## GDP GDP 0.47047915  
## Measles Measles 0.34343040  
## Status Status 0.02444965

boost\_model

## gbm(formula = Life\_exp ~ ., distribution = "gaussian", data = train,   
## n.trees = 500, interaction.depth = 4)  
## A gradient boosted model with gaussian loss function.  
## 500 iterations were performed.  
## There were 18 predictors of which 18 had non-zero influence.

yhat\_boost\_tr = predict (boost\_model , newdata=train)

## Using 500 trees...

mean((yhat\_boost\_tr -train$Life\_exp)^2)

## [1] 1.468854

train\_error=boost\_model$train.error[500]  
train\_error

## [1] 1.468854

yhat\_boost = predict (boost\_model , newdata=test)

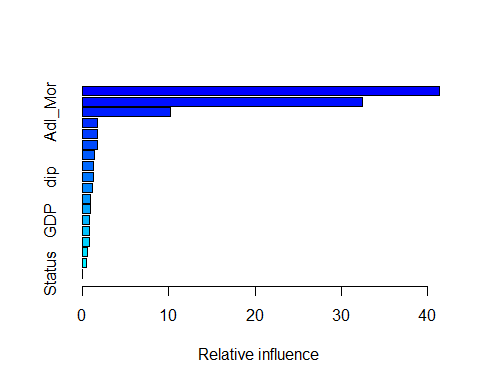
## Using 500 trees...

mean((yhat\_boost -test$Life\_exp)^2)

## [1] 3.359913

RESULT: Here we can see that the MSE is greater(3.35) than what we observed in the Random Forest Model, but intuitively this model should perform better than Random forest hence we tune the hyperparameters used in regression to find a better result.

boost\_model\_1=gbm(Life\_exp∼.,data=train, distribution="gaussian",n.trees=1000, interaction.depth=4,shrinkage=0.2)  
summary(boost\_model\_1)



## var rel.inf  
## Icm\_comp\_res Icm\_comp\_res 41.35581543  
## HIV HIV 32.45409608  
## Adl\_Mor Adl\_Mor 10.21065806  
## thin\_5\_9 thin\_5\_9 1.76481951  
## U5\_dth U5\_dth 1.76248831  
## schooling schooling 1.75960052  
## Alcohol Alcohol 1.44258136  
## Tot\_Exp Tot\_Exp 1.32186393  
## dip dip 1.27518106  
## POP POP 1.18708898  
## thin\_1\_19 thin\_1\_19 0.95474388  
## Polio Polio 0.92865016  
## GDP GDP 0.87244744  
## Per\_exp Per\_exp 0.79776843  
## BMI BMI 0.79631065  
## Measles Measles 0.56477023  
## Inf\_Dth Inf\_Dth 0.51384794  
## Status Status 0.03726803

boost\_model\_1

## gbm(formula = Life\_exp ~ ., distribution = "gaussian", data = train,   
## n.trees = 1000, interaction.depth = 4, shrinkage = 0.2)  
## A gradient boosted model with gaussian loss function.  
## 1000 iterations were performed.  
## There were 18 predictors of which 18 had non-zero influence.

yhat\_boost\_tr\_1 = predict (boost\_model\_1 , newdata=train)

## Using 1000 trees...

mean((yhat\_boost\_tr\_1 -train$Life\_exp)^2)

## [1] 0.2957571

train\_error=boost\_model\_1$train.error[500]  
train\_error

## [1] 0.8201739

yhat\_boost\_1 = predict (boost\_model\_1 , newdata=test)

## Using 1000 trees...

mean((yhat\_boost\_1 -test$Life\_exp)^2)

## [1] 3.303336

RESULT: Here we still see a MSE of 3.30 which is greater than the random forest model, hence we try another boosting technique in order to minimize our MSE.

*XGBoost Regression*: XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

library(xgboost)

##   
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':  
##   
## slice

library(caret)  
  
#define predictor and response variables in training set  
train\_x = data.matrix(train[, -2])  
train\_y = train[,2]  
  
#define predictor and response variables in testing set  
test\_x = data.matrix(test[, -2])  
test\_y = test[, 2]  
  
#define final training and testing sets  
xgb\_train = xgb.DMatrix(data = train\_x, label = train\_y)  
xgb\_test = xgb.DMatrix(data = test\_x, label = test\_y)  
  
#define watchlist  
watchlist = list(train=xgb\_train, test=xgb\_test)  
  
#fit XGBoost model and display training and testing data at each round  
model = xgb.train(data = xgb\_train, max.depth = 3, watchlist=watchlist, nrounds = 150)

## [1] train-rmse:48.885866 test-rmse:48.599966   
## [2] train-rmse:34.358407 test-rmse:34.132701   
## [3] train-rmse:24.197606 test-rmse:23.992596   
## [4] train-rmse:17.106676 test-rmse:16.909163   
## [5] train-rmse:12.162869 test-rmse:11.993329   
## [6] train-rmse:8.757393 test-rmse:8.602162   
## [7] train-rmse:6.423746 test-rmse:6.287008   
## [8] train-rmse:4.850176 test-rmse:4.720587   
## [9] train-rmse:3.844659 test-rmse:3.731270   
## [10] train-rmse:3.197061 test-rmse:3.131872   
## [11] train-rmse:2.816005 test-rmse:2.773826   
## [12] train-rmse:2.599887 test-rmse:2.589817   
## [13] train-rmse:2.485306 test-rmse:2.496112   
## [14] train-rmse:2.409794 test-rmse:2.438071   
## [15] train-rmse:2.349650 test-rmse:2.400877   
## [16] train-rmse:2.311196 test-rmse:2.367534   
## [17] train-rmse:2.283233 test-rmse:2.358001   
## [18] train-rmse:2.262698 test-rmse:2.348615   
## [19] train-rmse:2.228161 test-rmse:2.336063   
## [20] train-rmse:2.189216 test-rmse:2.291849   
## [21] train-rmse:2.179963 test-rmse:2.288818   
## [22] train-rmse:2.146972 test-rmse:2.259405   
## [23] train-rmse:2.103821 test-rmse:2.226035   
## [24] train-rmse:2.082421 test-rmse:2.211272   
## [25] train-rmse:2.062700 test-rmse:2.195364   
## [26] train-rmse:2.048640 test-rmse:2.189957   
## [27] train-rmse:2.022191 test-rmse:2.170071   
## [28] train-rmse:2.009406 test-rmse:2.167506   
## [29] train-rmse:1.986926 test-rmse:2.161757   
## [30] train-rmse:1.966040 test-rmse:2.148258   
## [31] train-rmse:1.957556 test-rmse:2.141874   
## [32] train-rmse:1.939965 test-rmse:2.129199   
## [33] train-rmse:1.929182 test-rmse:2.121249   
## [34] train-rmse:1.908607 test-rmse:2.112292   
## [35] train-rmse:1.889914 test-rmse:2.108607   
## [36] train-rmse:1.883767 test-rmse:2.102517   
## [37] train-rmse:1.863626 test-rmse:2.087322   
## [38] train-rmse:1.846863 test-rmse:2.076132   
## [39] train-rmse:1.839546 test-rmse:2.076130   
## [40] train-rmse:1.835969 test-rmse:2.076309   
## [41] train-rmse:1.827133 test-rmse:2.074057   
## [42] train-rmse:1.821710 test-rmse:2.071094   
## [43] train-rmse:1.810091 test-rmse:2.069146   
## [44] train-rmse:1.802887 test-rmse:2.070571   
## [45] train-rmse:1.781177 test-rmse:2.060202   
## [46] train-rmse:1.768772 test-rmse:2.055770   
## [47] train-rmse:1.755843 test-rmse:2.043726   
## [48] train-rmse:1.737845 test-rmse:2.036325   
## [49] train-rmse:1.729039 test-rmse:2.035825   
## [50] train-rmse:1.715042 test-rmse:2.034363   
## [51] train-rmse:1.702226 test-rmse:2.030374   
## [52] train-rmse:1.688104 test-rmse:2.029715   
## [53] train-rmse:1.677755 test-rmse:2.031180   
## [54] train-rmse:1.673617 test-rmse:2.030514   
## [55] train-rmse:1.667257 test-rmse:2.032290   
## [56] train-rmse:1.662944 test-rmse:2.031217   
## [57] train-rmse:1.647712 test-rmse:2.018569   
## [58] train-rmse:1.634009 test-rmse:2.004070   
## [59] train-rmse:1.628173 test-rmse:2.000250   
## [60] train-rmse:1.621742 test-rmse:2.000766   
## [61] train-rmse:1.611185 test-rmse:2.004085   
## [62] train-rmse:1.601974 test-rmse:2.001644   
## [63] train-rmse:1.591341 test-rmse:1.998494   
## [64] train-rmse:1.578281 test-rmse:1.990509   
## [65] train-rmse:1.563522 test-rmse:1.983234   
## [66] train-rmse:1.558212 test-rmse:1.983429   
## [67] train-rmse:1.555379 test-rmse:1.982659   
## [68] train-rmse:1.551635 test-rmse:1.979391   
## [69] train-rmse:1.547997 test-rmse:1.981130   
## [70] train-rmse:1.543175 test-rmse:1.978949   
## [71] train-rmse:1.535487 test-rmse:1.981520   
## [72] train-rmse:1.526488 test-rmse:1.977708   
## [73] train-rmse:1.519186 test-rmse:1.974122   
## [74] train-rmse:1.514526 test-rmse:1.975726   
## [75] train-rmse:1.510261 test-rmse:1.977669   
## [76] train-rmse:1.508342 test-rmse:1.976634   
## [77] train-rmse:1.489618 test-rmse:1.966832   
## [78] train-rmse:1.482319 test-rmse:1.965868   
## [79] train-rmse:1.469772 test-rmse:1.968201   
## [80] train-rmse:1.460227 test-rmse:1.958760   
## [81] train-rmse:1.456177 test-rmse:1.957099   
## [82] train-rmse:1.445713 test-rmse:1.960846   
## [83] train-rmse:1.437512 test-rmse:1.954412   
## [84] train-rmse:1.428553 test-rmse:1.948748   
## [85] train-rmse:1.426400 test-rmse:1.949385   
## [86] train-rmse:1.422236 test-rmse:1.949146   
## [87] train-rmse:1.416578 test-rmse:1.953630   
## [88] train-rmse:1.411582 test-rmse:1.952801   
## [89] train-rmse:1.404840 test-rmse:1.951736   
## [90] train-rmse:1.396928 test-rmse:1.952245   
## [91] train-rmse:1.385583 test-rmse:1.949514   
## [92] train-rmse:1.376815 test-rmse:1.946336   
## [93] train-rmse:1.365961 test-rmse:1.948457   
## [94] train-rmse:1.355987 test-rmse:1.942627   
## [95] train-rmse:1.348198 test-rmse:1.944429   
## [96] train-rmse:1.338418 test-rmse:1.947326   
## [97] train-rmse:1.332213 test-rmse:1.945586   
## [98] train-rmse:1.327774 test-rmse:1.944683   
## [99] train-rmse:1.319378 test-rmse:1.940492   
## [100] train-rmse:1.308355 test-rmse:1.925803   
## [101] train-rmse:1.300122 test-rmse:1.925992   
## [102] train-rmse:1.293171 test-rmse:1.920970   
## [103] train-rmse:1.288561 test-rmse:1.922004   
## [104] train-rmse:1.279105 test-rmse:1.912979   
## [105] train-rmse:1.272277 test-rmse:1.909133   
## [106] train-rmse:1.270822 test-rmse:1.909632   
## [107] train-rmse:1.261751 test-rmse:1.906796   
## [108] train-rmse:1.254993 test-rmse:1.908864   
## [109] train-rmse:1.245469 test-rmse:1.907942   
## [110] train-rmse:1.236430 test-rmse:1.906554   
## [111] train-rmse:1.230357 test-rmse:1.898945   
## [112] train-rmse:1.223959 test-rmse:1.896748   
## [113] train-rmse:1.221312 test-rmse:1.897170   
## [114] train-rmse:1.212456 test-rmse:1.894410   
## [115] train-rmse:1.204239 test-rmse:1.892122   
## [116] train-rmse:1.196370 test-rmse:1.890680   
## [117] train-rmse:1.190156 test-rmse:1.887716   
## [118] train-rmse:1.183643 test-rmse:1.889520   
## [119] train-rmse:1.181259 test-rmse:1.891334   
## [120] train-rmse:1.179589 test-rmse:1.889942   
## [121] train-rmse:1.173458 test-rmse:1.891514   
## [122] train-rmse:1.166686 test-rmse:1.893957   
## [123] train-rmse:1.159762 test-rmse:1.898121   
## [124] train-rmse:1.154990 test-rmse:1.895229   
## [125] train-rmse:1.147294 test-rmse:1.893168   
## [126] train-rmse:1.145295 test-rmse:1.891431   
## [127] train-rmse:1.142767 test-rmse:1.889913   
## [128] train-rmse:1.138075 test-rmse:1.889326   
## [129] train-rmse:1.130493 test-rmse:1.891842   
## [130] train-rmse:1.125882 test-rmse:1.889387   
## [131] train-rmse:1.120514 test-rmse:1.888057   
## [132] train-rmse:1.114764 test-rmse:1.888842   
## [133] train-rmse:1.109641 test-rmse:1.891384   
## [134] train-rmse:1.105117 test-rmse:1.891109   
## [135] train-rmse:1.102306 test-rmse:1.894137   
## [136] train-rmse:1.101531 test-rmse:1.893908   
## [137] train-rmse:1.097329 test-rmse:1.893561   
## [138] train-rmse:1.093667 test-rmse:1.893457   
## [139] train-rmse:1.091884 test-rmse:1.892819   
## [140] train-rmse:1.086791 test-rmse:1.891226   
## [141] train-rmse:1.082955 test-rmse:1.893098   
## [142] train-rmse:1.081276 test-rmse:1.893297   
## [143] train-rmse:1.080083 test-rmse:1.893350   
## [144] train-rmse:1.079159 test-rmse:1.891123   
## [145] train-rmse:1.074135 test-rmse:1.888912   
## [146] train-rmse:1.072416 test-rmse:1.888849   
## [147] train-rmse:1.065373 test-rmse:1.890415   
## [148] train-rmse:1.061411 test-rmse:1.890137   
## [149] train-rmse:1.056571 test-rmse:1.886697   
## [150] train-rmse:1.054173 test-rmse:1.886697

#define final model  
final\_xgb = xgboost(data = xgb\_train, max.depth = 3, nrounds = 117, verbose = 0)  
final\_xgb

## ##### xgb.Booster  
## raw: 134.3 Kb   
## call:  
## xgb.train(params = params, data = dtrain, nrounds = nrounds,   
## watchlist = watchlist, verbose = verbose, print\_every\_n = print\_every\_n,   
## early\_stopping\_rounds = early\_stopping\_rounds, maximize = maximize,   
## save\_period = save\_period, save\_name = save\_name, xgb\_model = xgb\_model,   
## callbacks = callbacks, max.depth = 3)  
## params (as set within xgb.train):  
## max\_depth = "3", validate\_parameters = "1"  
## xgb.attributes:  
## niter  
## callbacks:  
## cb.evaluation.log()  
## # of features: 18   
## niter: 117  
## nfeatures : 18   
## evaluation\_log:  
## iter train\_rmse  
## 1 48.885866  
## 2 34.358407  
## ---   
## 116 1.196370  
## 117 1.190156

pred\_y = predict(final\_xgb,xgb\_test)  
mean((test\_y - pred\_y)^2)

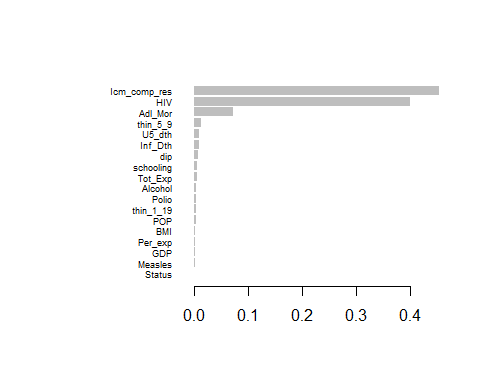
## [1] 3.563473

RESULT: Here we get a MSE of 3.56 which is still greater than the random Forest Model.

xgb\_imp<-xgb.importance(model=final\_xgb)  
xgb\_imp

## Feature Gain Cover Frequency  
## 1: Icm\_comp\_res 0.4536708863 0.133104206 0.103359173  
## 2: HIV 0.3994452441 0.088096775 0.073643411  
## 3: Adl\_Mor 0.0721553185 0.141239375 0.151162791  
## 4: thin\_5\_9 0.0138294814 0.061937774 0.056847545  
## 5: U5\_dth 0.0094807458 0.030431920 0.025839793  
## 6: Inf\_Dth 0.0085284444 0.027668049 0.036175711  
## 7: dip 0.0081880323 0.051191920 0.040051680  
## 8: schooling 0.0059325627 0.049216794 0.054263566  
## 9: Tot\_Exp 0.0054065186 0.068707303 0.074935401  
## 10: Alcohol 0.0046044052 0.050857843 0.056847545  
## 11: Polio 0.0040083378 0.024993807 0.033591731  
## 12: thin\_1\_19 0.0030859922 0.038852276 0.041343669  
## 13: POP 0.0029894286 0.061823699 0.060723514  
## 14: BMI 0.0023422199 0.027304639 0.051679587  
## 15: Per\_exp 0.0022302218 0.038775683 0.037467700  
## 16: GDP 0.0022165677 0.059382496 0.056847545  
## 17: Measles 0.0015925397 0.044119277 0.042635659  
## 18: Status 0.0002930528 0.002296164 0.002583979

xgb.plot.importance(xgb\_imp)



1. ***INTERPRETATION AND ANALYSIS***

Our main objective for the given analysis was to build an efficient and accurate regression model in order to predict our response Variable that is “Life Expectancy” given the predictor variables present in the data. We initially cleaned the dataset and removed and imputed the missing values present. Then we ran test in order to identify any outliers or Multicollinearity present in the dataset. After getting all the above mentioned information we started with our regression analysis.

We first used the simple regression models like Ordinary Least Squares where we saw only 83.5% of our response variable was being explained by our predictors and there was also 2 pairs of predictors which had multicollinearity which were “ thinness 5-9” - “thinness 1-19”

And “infant mortality” – “Under 5 Deaths”.

We then decided to move forward in the direction of Regularization and Feature Selection Techniques like LASSO, Ridge and Stepwise Method in order to eliminate these problems and built regression models on them after splitting the dataset in Train and Test with a 7:3 ratio but as described above we saw that our values of Root mean squared errors were quite high on the test dataset ranging between 10-12.

Consequently, we decide to move towards advanced regression techniques like Bagging and Boosting. Hence we used random forest, Gradient Boosting and XGBoosting to run regression. Here gladly we saw that the value of root mean squared values ranged from 1.7-1.9 on the test data. Here the best result was seen in the Random Forest Regression Model which had a Root mean squared value of error equal to 1.701.

Hence, we can conclude our analysis by saying that using the Random Forest regression model we can predict the life expectancy of any country, developing or developed, given that we have the data for all the predictor variables, within an error margin of 1.7 years.

1. ***REFERENCES*** 
   * 1. [ISLR .pdf: Machine learning masters with deep learning and internship (instructure.com)](https://canvas.instructure.com/courses/2412303/files/114118826?module_item_id=38748066)
     2. <https://stackoverflow.com/>
     3. <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>
     4. [What is Random Forest? | IBM](https://www.ibm.com/cloud/learn/random-forest)
     5. [XGBoost in R: A Step-by-Step Example (statology.org)](https://www.statology.org/xgboost-in-r/)
     6. [randomForest function - RDocumentation](https://www.rdocumentation.org/packages/randomForest/versions/4.7-1/topics/randomForest)
     7. [Practical Tutorial on Random Forest and Parameter Tuning in R Tutorials & Notes | Machine Learning | HackerEarth](https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/)
     8. [Tree-Based Models in R (slcladal.github.io)](https://slcladal.github.io/tree.html)