JSC370 Final Project

Shiyuan Zhou

2022/4/12

Introduction

How humans can live longer is one of the most debated topics in human history. In an era of rapid advances in medicine, education, and technology, human health has improved significantly but will humans necessarily live longer? The differences between countries are not only due to race and region but the imbalance on health care and medical technologies across countries. However, in many cases, it is difficult to spread or teach advanced medical treatments to other countries. Does the absence of them determines lower average life expectancy? Obviously not, medical treatment is not the only factor that determines the life expectancy of human beings but also the climate and environment, social factors, etc.

In fact, most of the influencing factors come from the government and social organizations. We have a certain measure of whether the government is making a difference in humanistic care, which is the Human Development Index (HDI). The HDI is defined as a summary measure of average achievement in key dimensions of human development, like health. Also, social care as well as health care development cannot be achieved without the government expenditure. That is, at the theoretical level, both the HDI and government expenditure on health care may have an impact on the health of people, leading to an increase in their average life expectancy. However, governments' decisions heavily depends on the level of development of the country, i.e., whether a country is a developed country or not also affects the health policy and the standard of living of the population. These thoughts have led me to wonder whether life expectancy has a stronger relationship between human development index or government health care spending. Additionally, whether these relationships will be altered by the degree of development.

However, HDI, health expenditure, and developing status may be not enough to predict life expectancy. There are other social factors that may have a big influence on local life expectancy, like adult mortality. Base on recorded social factors observations, we can add them into more complex models and see how life expectancy would be in the future. By having that high-dimensional model, we would help to increase life expectancy. The whole society will be benefited as more and more related social organizations can get involved and be improved. Hence, we could generalize our new question: how to accurately predict life expectancy?

Git-hub repo: https://github.com/ZhouEEEEEE/JSC370-Final-Project

Research Question

The aim for this whole project is to increase life expectancy. As our interest in HDI and health expenditure, we have our first two research questions to see possible positive relationship. Additionally, we could use our built model in the third question to give better prediction on life expectancy.

1. Is government health expenditure have greater impact on life expectancy than Human Development Index?
2. Does life expectancy also depends on the development status of the country?
3. How to accurately predict life expectancy by social factors?

Methods

Used R packages

Here are the following R packages that I used for this portfolio.: data.table; dtplyr; dplyr; ggplot2; mgcv; zoo; leaflet; ggpubr; lme4; lmtest; tidyverse; rpart; rpart.plot; randomForest; gbm; xgboost; caret

Data Source

The Data that I used to answer my research question is based on the WHO data and published on Kaggle by Kumar Rajarshi. This dataset includes values social factors of 193 countries from 2000 to 2015 and the life expectancy in age. In our research question, we are aim to compare the impact of government health expenditure and Human Development Index on life expectancy. These two predictors are represent by 'Total expenditure' and 'Income composition of resources' in our data-set. For our third research question, the target is life expectancy. Since we also stated that social factors may have a big difference between developed and developing countries. We sill also include the binary variable 'Status' that indicate the development status of a country. All of these variables will change across years.

Link of data: https://www.kaggle.com/kumarajarshi/life-expectancy-who

We also used a data-set that help us in visualizations from: https://www.kaggle.com/datasets/andradaolteanu/country-mapping-iso-continent-region?resource=download. This is not our main data and all the data exploration part is focusing on our previous main data.

Exploratory Data Analysis

Before answering our research question, we need to do Exploratory Data Analysis first to find issues in our data, clean our data, and make summary statistics, plots, and graphs for our key variables.

Data Checking

Table 1: Number of missing values for each variable

	num_na
Country	0
Year	0
Status	0
Life expectancy	10
Adult Mortality	10
infant deaths	0
Alcohol	193
percentage expenditure	0
Hepatitis B	553
Measles	0
BMI	34
under-five deaths	0
Polio	19
Total expenditure	225
Diphtheria	19
HIV/AIDS	0
GDP	448

	num_na
Population	652
thinness 1-19 years	34
thinness 5-9 years	34
Income composition of resources	167
Schooling	163

We have 2937 number of observations and 22 number of variables in our dataset. There are 14 columns contain NAs. There are 167 missing values in Income composition of resources and 225 NAs in total expenditure. The variable with highest amount of NAs is 'Hepatitis B'. We will do the missing value imputation in the next section.

Check dimensions of our data

Table 2: Summery table of the dimensions of our data

axis	value
num_observations	2937
num_variables	22

We have 2937 number of observations and 22 number of variables in our dataset.

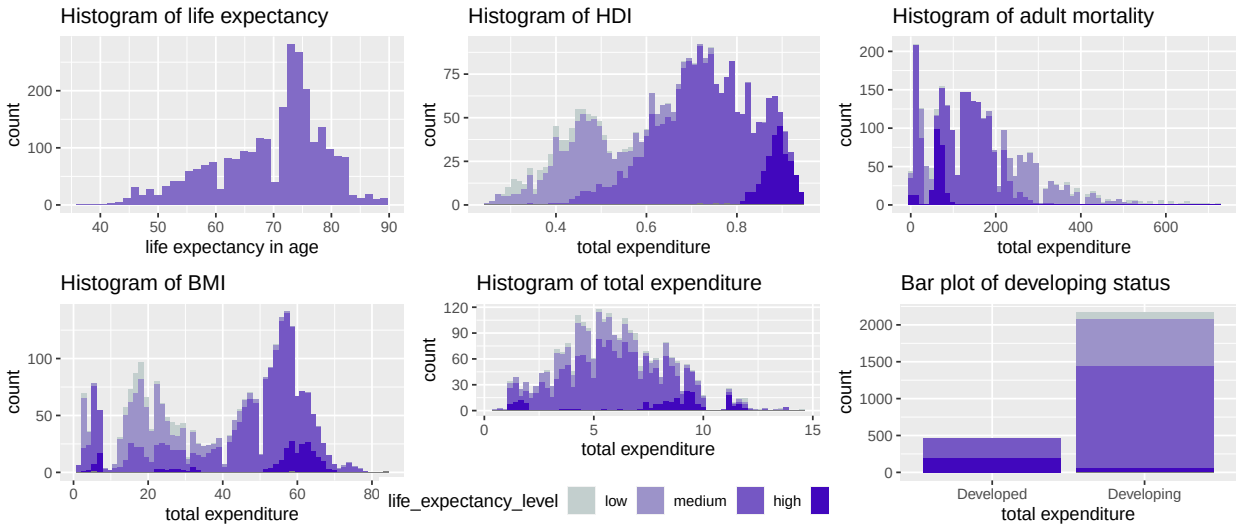
Check the summary statistics of required numeric variables

Table 3: Summary statistics of required variables

Life expectancy	Adult Mortality	Total expenditure	HIV/AIDS	Income composition of resources
Min. :36.30	Min. : 1.0	Min. : 0.370	Min. : 0.100	Min. :0.0000
1st Qu.:63.10	1st Qu.: 74.0	1st Qu.: 4.260	1st Qu.: 0.100	1st Qu.:0.4930
Median :72.10	Median :144.0	Median : 5.755	Median : 0.100	Median :0.6770
Mean :69.22	Mean :164.8	Mean : 5.938	Mean : 1.743	Mean :0.6275
3rd Qu.:75.70	3rd Qu.:228.0	3rd Qu.: 7.492	3rd Qu.: 0.800	3rd Qu.:0.7790
Max. :89.00	Max. :723.0	Max. :17.600	Max. :50.600	Max. :0.9480
NA's :10	NA's :10	NA's :225	NA	NA's :167

Since our data-set contains multiple variables, presenting summary statistics for all the variables is not optimal. Here are the summary statistics of several key variables help us to find the issues and reliability of our data. According to the summary table we get, variable 'Life expectancy' and 'Total expenditure' do not have big issues and in our estimated bound(life expectancy should be greater than 0 and less than 100, total expenditure should be greater than 0 and less than 100 since it represents proportion). However, the variable 'income composition of resources' has minimum values equals to 0. Since this variable indicate human development index, its impossible to have 0 values, which means we need to remove those observations. According to the worldpopulationreview.com, the country with lowest HDI in 2019 is Niger with 0.394. Hence, 0 income composition should be removed from the data set in order to prevent wrong model fitting. Other variables' reliability were also checked.

Check Distribution of required variables The distribution of several main variables are checked base on their plotted histograms.



To have more insights on their relationship with life expectancy, we also made variable 'life_expectancy_level' for different levels of ages only for EDA. The distribution of life expectancy and total expenditure is quite normal but life expectancy is left-skewed. For HDI(Income composition of resources) and BMI, we can see that, higher level of life expectancy concentrated on higher HDI, which may indicate a positive relationship. For adult mortality records, low mortality may have higher life expectancy as the distribution of color goes darker from right to left. FUrthermore, developed country tend to have higher life expectancy. There were no too much clear relationship in other variables.

Data Wranggling

NA Imputation and data joining We handled the missing values by imputation. We use mean value of current column to impute by for looping each column. In the visualization part, we also used countries' sub region of their continent. Hence, we made a left join on our main data-set with continent data-set by each country's name.

Create New Variable To do further data exploration on different types of plots, we need both numeric and categorical 'Total expenditure' and 'Income composition of resources'. Converting current numeric variables to categorical variables helps us on stacked histograms, statistical summary graph, and etc. In many statistical research on social factors, health expenditure and HDI are always represented by different levels.

Create a new categorical variable named "expenditure_level" using total expenditure on health of a country. (rare total expenditure < 3; low total expenditure 3-5; mild total expenditure 5-9; high total expenditure $>$ 9) and a new categorical variable named "hdi_level" indicating level of income composition of resources of countries(low income composition < 0.55; medium income composition 0.55-0.7; high income composition 0.7-0.8; very high income composition > 0.8). Additionally, we should use factor() function to give our levels an order for future convenience.

Since we also need to perform gradient boosting and extreme gradient boosting to predict life expectancy based on our dataset. Hence, we need to make character variables into numeric variables since boosting model cannot apply to categorical variables. By looking at the dataset, we found that variable 'Status' and 'Country' are categorical. We do not need variable 'Country' in machine learning model fitting as we investigate the dataset as a whole: each country's value in each year is a single observation. Variable 'Status' is binary. Hence, we only need to convert it into 1 and 0 and create new variables 'status_num'.

We also found the range of variable 'GDP', 'Percentage Expenditure' and 'Population' are much larger than other variables, which means we need to scale them. If there is a big difference in the range of variables, higher ranging numbers may have superiority in model fitting.

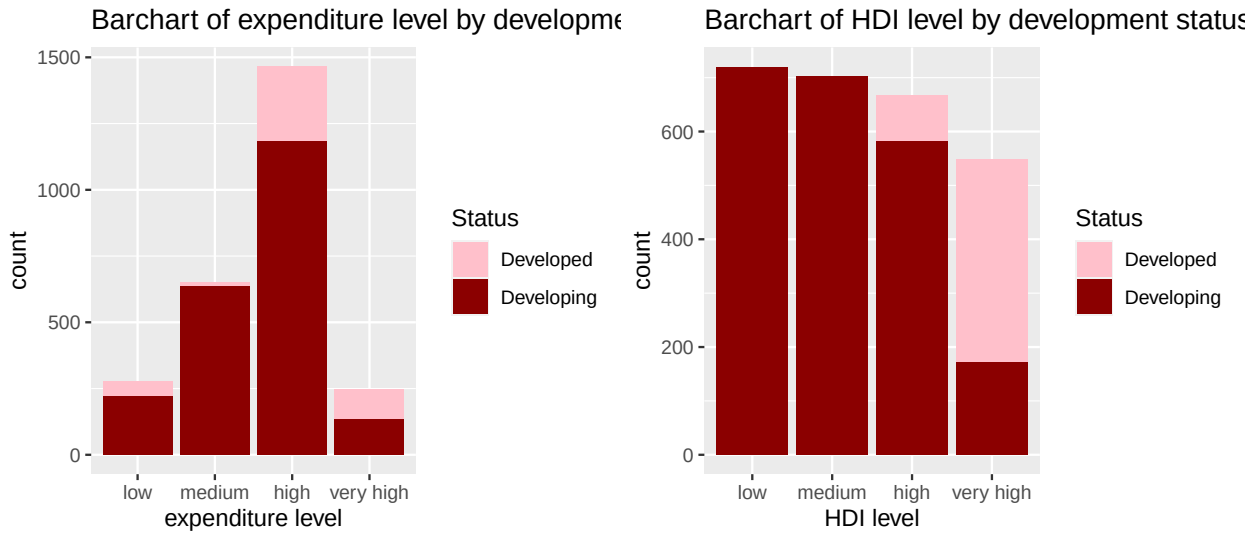
Visualizations

The interactive version of histograms, bar charts, and statistical summary graph is presented in the 'Visualization' page on the website of this project: https://zhoueeeeee.github.io/JSC370-Final-Project/.

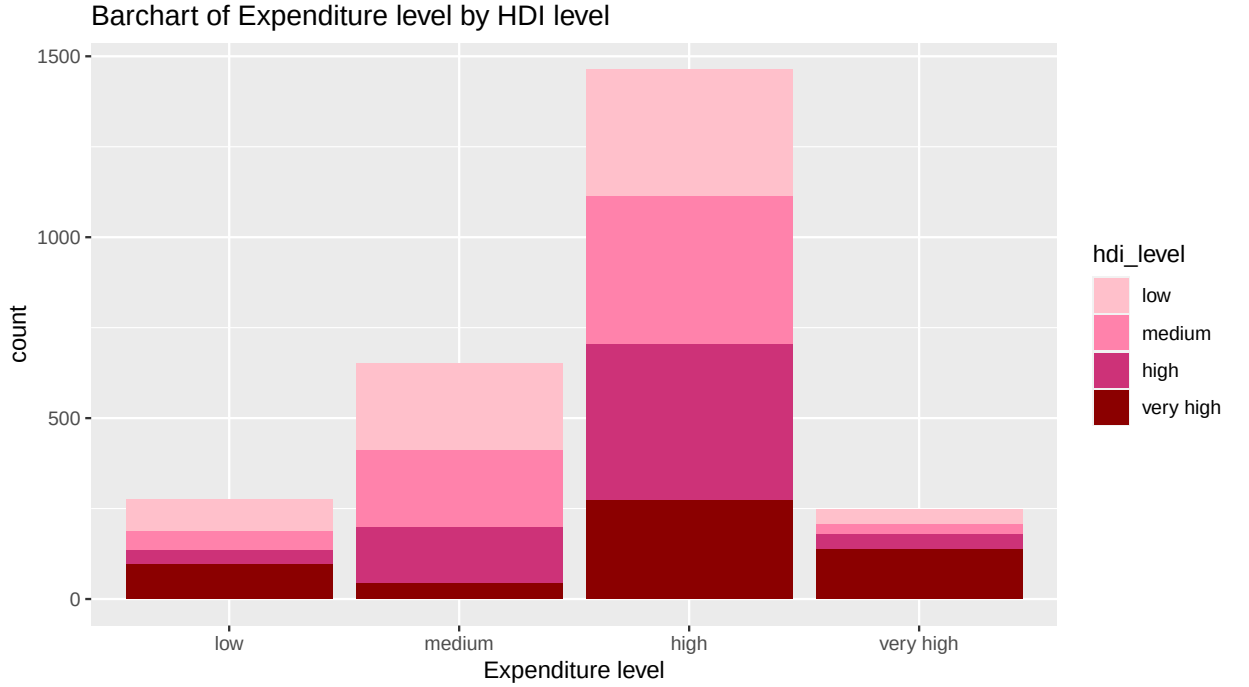
Histograms

The first plot we have is the stacked histograms of life expectancy by expenditure level and HDI level. The proportion of each level of expenditure does not make a big difference across different ages. However, in the stacked histogram for HDI level, it is very clear that higher HDI level become more concentrated on the right, which is higher age. For different range of age, there always have a dominated HDI level. For example, for life expectancy less than 60, low HDI level dominates. Hence, according to these histograms, income composition of resources have a stronger relationship with life expectancy.

Bar Chart

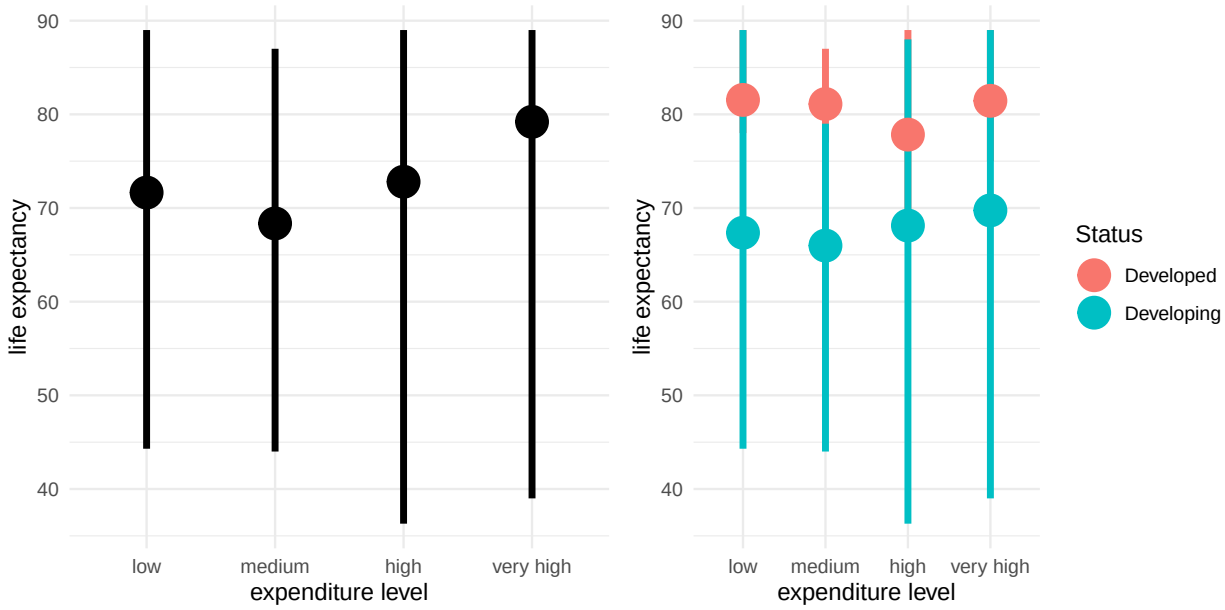


According to the barcharts we have for categorical variables expenditure level and HDI level by development status, developed countries tend to have higher expenditure and income composition of resources. However, the trend between HDI level and development status is stronger.

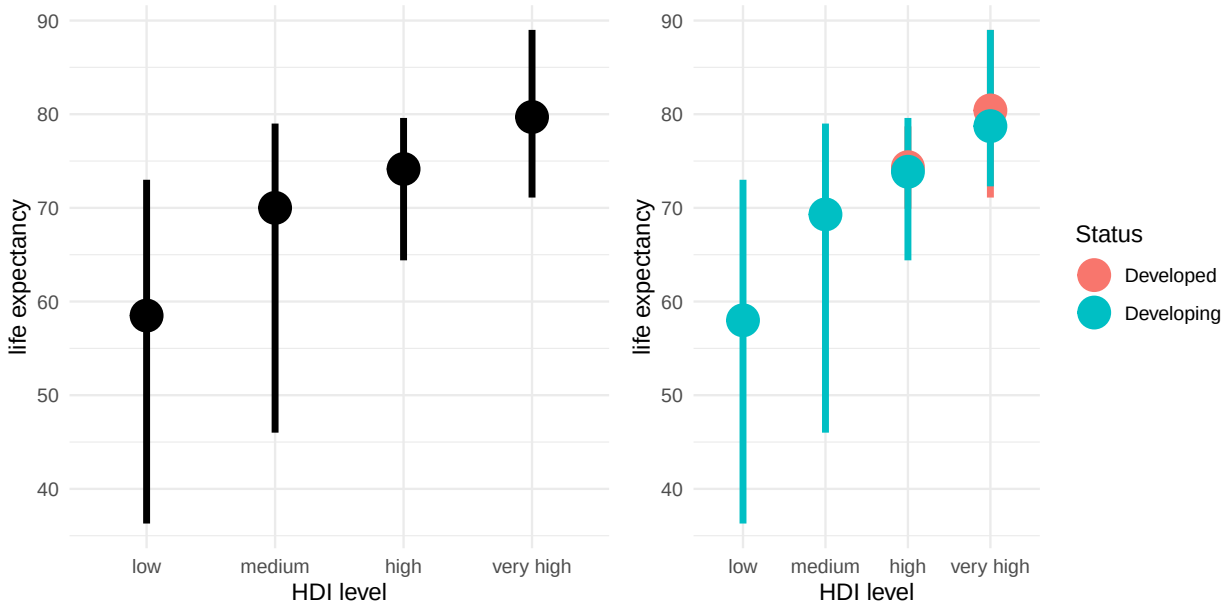


According to the bar chart we get for expenditure level by HDI level, the proportion of low and high HDI level is the largest in the low expenditure level. The proportion of medium HDI level is the largest in medium expenditure level and also for high and very high level. Hence, we could say counties with high expenditure level have higher probability to have high HDI level. Predictors expenditure and HDI may have a linear relationship.

Statistical summary graph



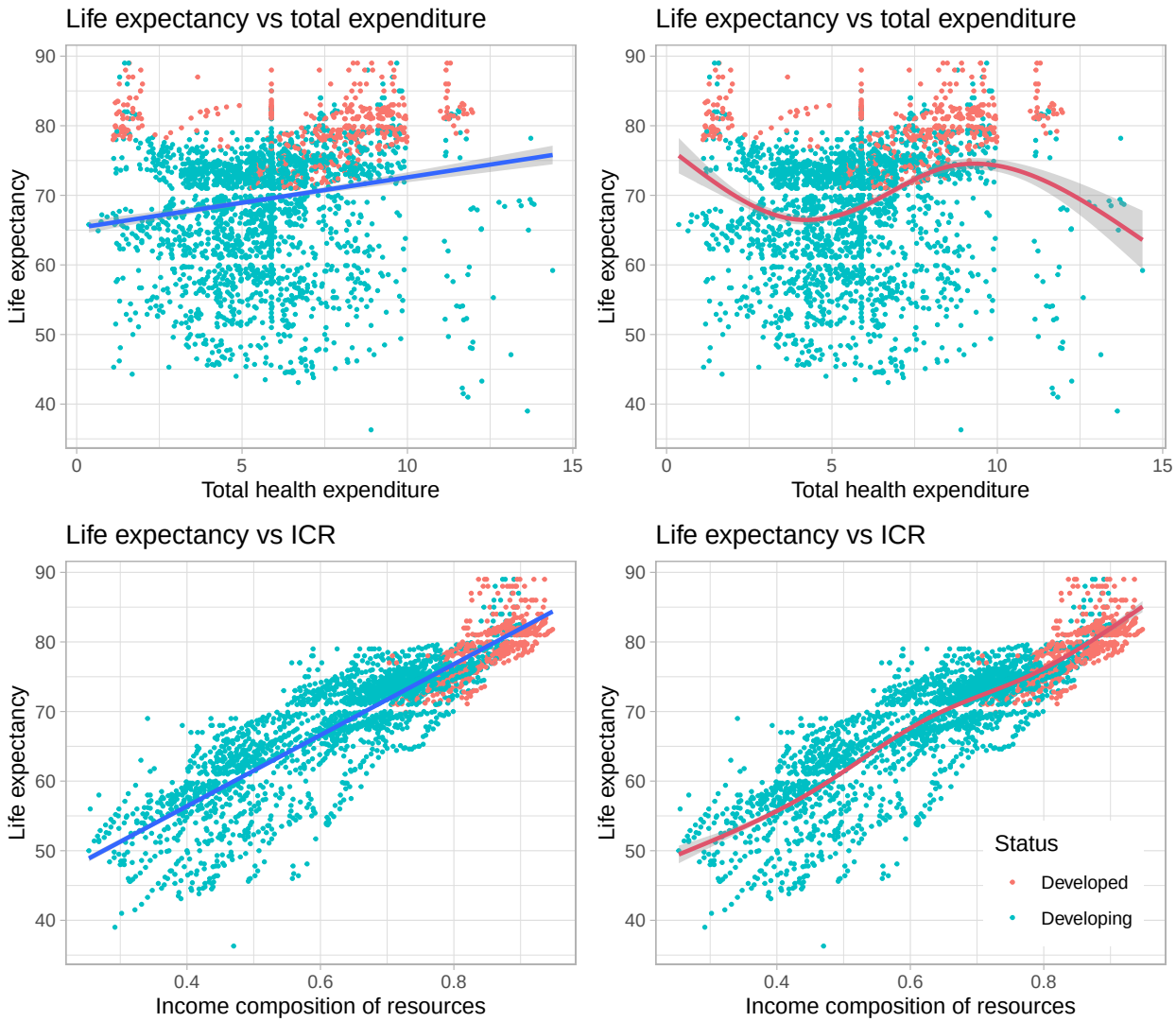
According to the statistical summary graph for expenditure level, though mean of life expectancy in low level of expenditure level is higher, we may have an increasing trend between expenditure level and life expectancy. However, if we adjusted by development status, we can see that the trend is clear for developing countries but not for developed countries. The higher mean of life expectancy in low level of expenditure was pulled up by the values of developed countries as the orange points shows. Additionally, the range of life expectancy for each expenditure level as the distance from min to max is large, which means our model may not fit tightly.



The statistical summary graph we have for HDI level shows a positive relationship between human development index and life expectancy. Adjusting by development status did not make a difference on our relationship. Additionally, the distance between min and max is much shorter than that in expenditure level which indicate a strong relationship and a tighter model fit.

Scatterplots

The scatter plot of live expectancy vs total health expenditure and income composition of resources clearly present what actual model fitting will be in our dataset.



The two plots with two blue straight line on the left is the linear model fitted in each of the relationship. The two plots on the right is the cubic spline model we have, where the red curve is the fitted splines. The middle straight dotted line in expenditure plot is where we impute the NAs in variable total expenditure with mean. According to the plots we have for live expectancy vs total health expenditure, the linear model is not very fitted to our data. Spline model could explain more variation and yields better fit but the decreasing trend when total health expenditure is greater than 12.5 may comes from over-fitting on the right-most points. Comparing to what we have in life expectancy vs income composition of resources, a positive linear trend is pretty clear. However, the fitted spline model does not make a big difference than the linear model. We need to further decide which model is better by adjusted R squared since spline model may have a higher adjusted R squared but the cost is over-fitting.

Scatter plot of life expectancy vs HIV/AIDS

This plot is shown in the 'Plot 1' under the interactive visualization section in page: https://zhoueeeeee.github.io/JSC370-Final-Project/

HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years) HDI level: Levels of income composition of resources of countries('low' income composition < 0.55; 'medium' income composition 0.55-0.7 'high' income composition 0.7-0.8; 'very high' income composition > 0.8).

According to our scatter plot for life expectancy versus HIV/AIDS deaths, we can see a inverse relationship that is quite poisson distributed. As the deaths goes higher, the life expectancy decreases and becomes flat when the HIV/AIDS is above 15. We did have several outliers that in the bottom of the plot that shows low HIV/AIDS but low life expectancy, which means there may be other influential factors in that observation lead to low ages. We also included HID-levels in the plot to group our observations. It shown that high HID level countries having high life expectancy and low HIV/AIDS deaths as the blue dots and purple dots concentrated on the left. Indicating a inverse relationship between HIV/AIDS and HDI and a positive relationship between Life expectancy and HDI exist.

Scatterplot of Life Expectancy vs Adult-Mortality for each level HDI level

This plot is shown in the 'Plot 2' under the interactive visualization section in page: https://zhoueeeeee.github.io/JSC370-Final-Project/

Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

According to the plot we have for Life Expectancy vs Adult-Mortality in 2013. We able to get insight in their relationship, which is a inverse linear relationship. Higher adult Mortality may result in low life expectancy. I grouped each dot(country) by their HDI level in 2013 and controlled the size of the dot by each county's total health expenditure in 2013. We can see that countries with relatively low adult mortality and high life expectancy tend to have higher health expenditure. Similar to previous scatter-plot result, high HDI countries have lower adult mortality rate. Additionally, according to dots' size. Countries with higher health expenditure tend to have lower adult mortality rate and high life expectancy. However, We also had country 'Lesotho' that spend a lot on health expenditure but failed to reduce adult mortality and increase life expectancy. We may use its lower HDI level to explain the situation.

Line Graph of Life Expectancy vs Income Composition of Resources for each Sub-region Group of Country

This plot is shown in the 'Plot 3-A' under the interactive visualization section in page: https://zhoueeeeee.github.io/JSC370-Final-Project/

In this line graph, we have a line for life expectancy vs income composition of resources for a country. To make the graph clearer, I grouped countries into different sub-region of their continents. We can investigate that, generally, higher income composition of resources result in higher life expectancy in most of the countries. Hence, we might conclude that there is a positive relationship between life expectancy and income composition of resources, which is Human Development Index by this graph. Furthermore, we could also see that region sub-Saharan Africa had low HDI and low life expectancy, which may due to their poverty issues. Regions in Europe and 'Australia and New Zealand' had pretty high HDI and high life expectancy. We removed countries 'Haiti', 'United Kingdom of Great Britain and Northern Ireland', 'United Republic of Tanzania', 'Cote d'Ivoire', and 'Republic of Korea' since their ICR values are imputed as they were missing in data collection.

Line Graph of Life Expectancy vs Total Expenditure for each Sub-region Group of Country

This plot is shown in the 'Plot 3-B' under the interactive visualization section in page: https://zhoueeeeee.github.io/JSC370-Final-Project/

To compare with the line graph we have for HDI, a line graph of total health expenditure versus life expectancy was also made. We cannot see any clear relationship between those two variables, which means health expenditure might be a less effective factor than HDI. Most of the regions having countries that across each level of expenditure. However, countries in sub-Saharan African also had lower life expectancy values.

Stacked Histogram of Schooling by Life Expectancy Levels

This plot is shown in the 'Plot 4' under the interactive visualization section in page: https://zhoueeeeee.github.io/JSC370-Final-Project/

Schooling: Number of years of Schooling(years)

The stacked histogram of variable 'Schooling' was made. We differed each bar by life expectancy level. According to the plot, countries with high schooling had higher life expectancy level as the color goes darker from left to right. Most of the countries achieve high level of life expectancy when they had more than 10 years of schooling. There is no really big difference between the schooling of the counties with 'low' and 'medium' life expectancy level. Hence, we may have a weak positive relationship between schooling and life expectancy.

Model Fitting

Inferential Models Comparing HDI and Health Expenditure

To compare whether HDI or health expenditure have stronger relationship with life expectancy, we use them as predictors and fit linear. linear mixed (year as random effect), and spline models. Since we also want to add consideration of development status, we will fit all the models and adjusted by status again.

a) Models without adjusted by development status

Linear models:

M1: Total expenditure as predictor: lm(life_exp ~ total_exp, data = ds)
M2: Income composition of resources as predictor: lm(life_exp ~ income_com, data = ds)

Linear mixed models:

M3: Total expenditure as fixed effect and year as random effect: lmer(life_exp ~ total_exp + (1|year), data = ds)
M4: Income composition of resources as fixed effect and year as random effect: lmer(life_exp ~ income_com + (1|year), data = ds)

Spline models:

M5: Total expenditure as smooth terms: gam(life_exp~ s(total_exp,bs="cr",k=3),data=ds)
M6: Income composition of resources as smooth terms: gam(life_exp~ s(income_com,bs="cr",k=3),data=ds)

b) Models with adjusted by development status

Linear models:
M7: Total expenditure and status as predictor: lm(life_exp ~ total_exp + status_ind, data = ds)
M8: Income composition of resources and status as predictor: lm(life_exp ~ income_com + status, data = ds)

Linear mixed models:

M9: Total expenditure and status as fixed effect and year as random effect: lmer(life_exp ~ total_exp + status + (1|year), data = ds)
M10: Income composition and status of resources as fixed effect and year as random effect: lmer(life_exp ~ income_com + status + (1|year), data = ds)

Spline models:
M11: Total expenditure as smooth terms adjusted by status: gam(life_exp~ s(total_exp, bs="cr",k=5) + status,data=ds)
M12: Income composition of resources as smooth terms adjusted by status: gam(life_exp~ s(income_com, bs="cr",k=3) + status, data=ds)

Machine Learning Models Predict on Life Expectancy

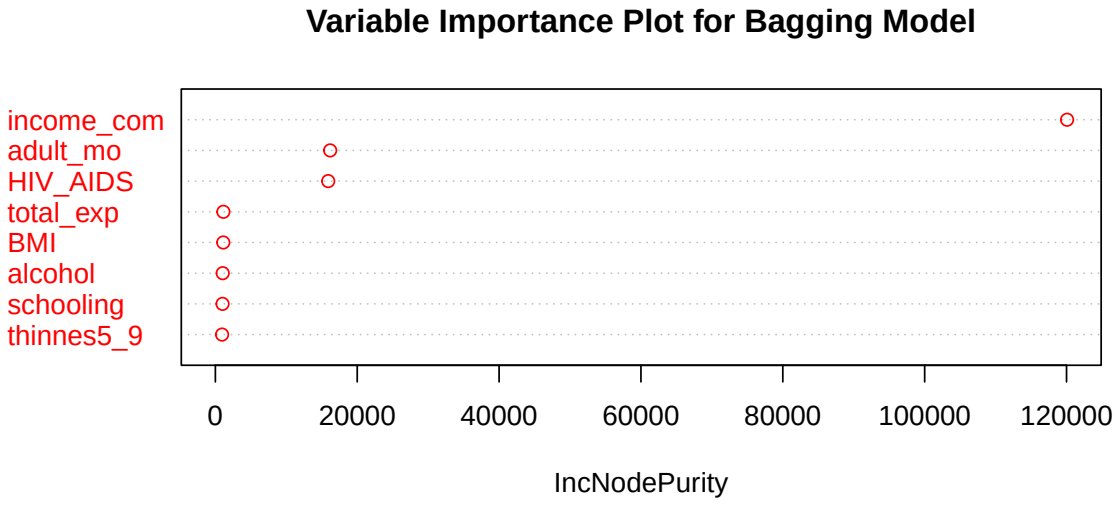
Discussed Variable meaning:

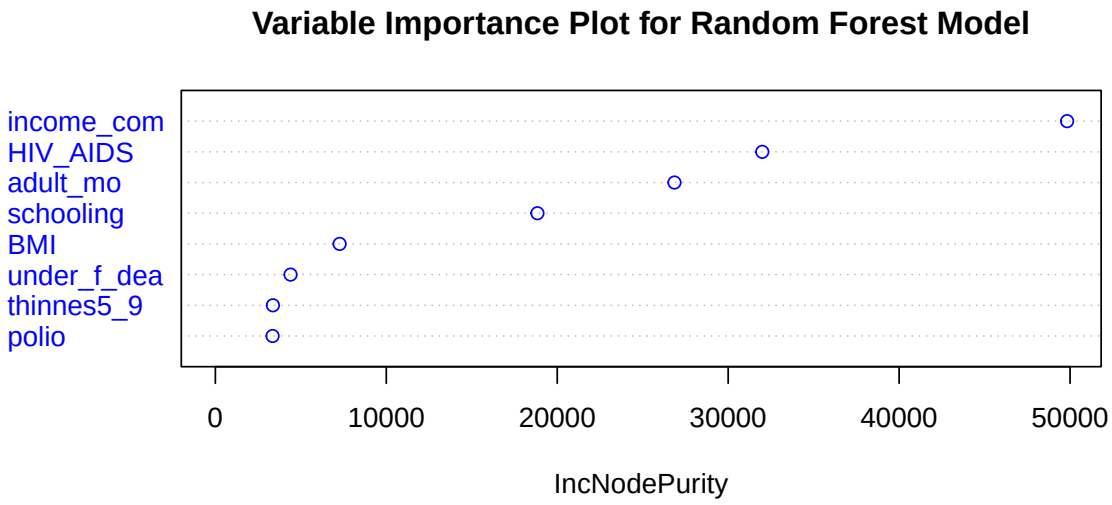
- income_com: Income composition of resources, which is HDI in our research.
- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)
- Schooling: Number of years of Schooling(years)
- BMI: Average Body Mass Index of entire population

Regression Tree In this section, we will perform basic machine learning techniques on our life expectancy data. We will fit regression tree, bagging, random forest, gradient boosting, and extreme gradient boosting models to predict life expectancy based on the social factors in the data. Our aim is finding a most predictive model from them by comparing their MSE.

By fitting a regression tree, we are able to find a optimal complexity parameter that has the minimal cross-validation error in the CP table. Based on that optimal complexity parameter, we are able to pruned the tree which help to reduce complexity and over-fitting of our decision tree model and improve prediction. Its MSE was also calculated for further model comparison. Since there are too may splits in our pruned regression tree. Visualizing it could be difficult but we can still evaluate it by its MSE.

Bagging and Radom Forest Fitting random forest and bagging models help us to find the most important features to predict life expectancy.

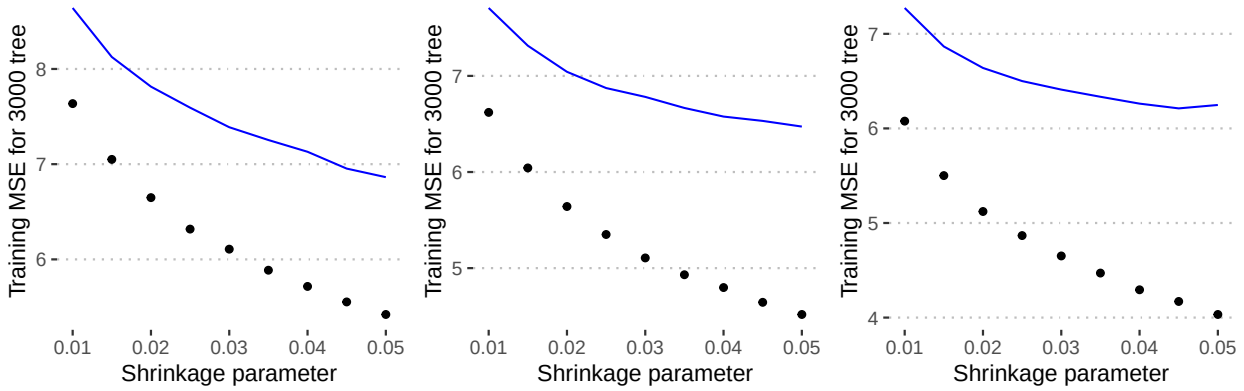




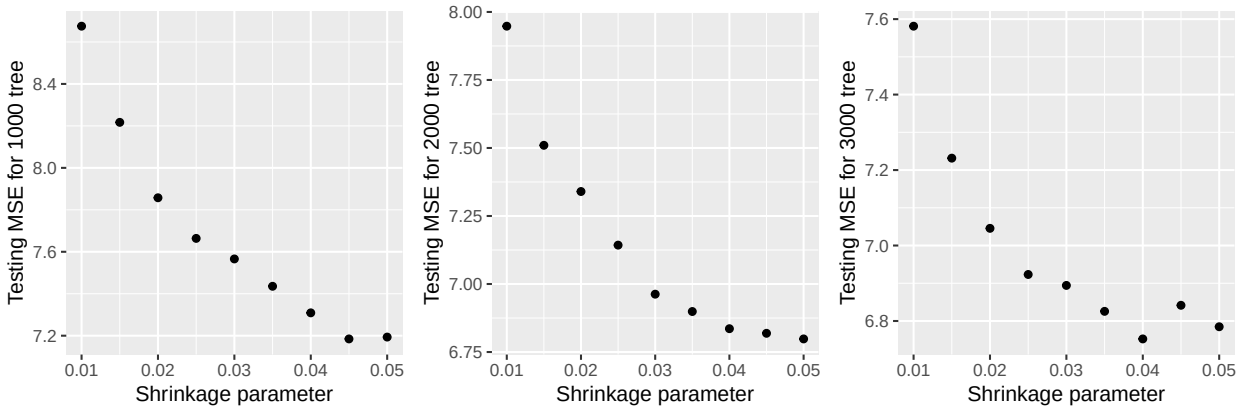
According to the variable importance plots we have for bagging and random forest. Variable 'income_com' is the most important features in both models. Adult mortality, HIV/AIDS deaths, and BMI are also quite significant in both models.

Gradient Boosting Model Gradient Boosting Model was also fitted to our data. To improve performance, we did parameter tuning on both shrinkage parameter and number of trees. We picked 1000, 2000, and 3000 as possible numbers of trees. Since there is pretty much tree fitted, which means a slightly large learning rate could be helpful on decaying the gradient. Hence, the range for possible learning rate we picked is 0.01 to 0.05 by 0.0005 on each step.

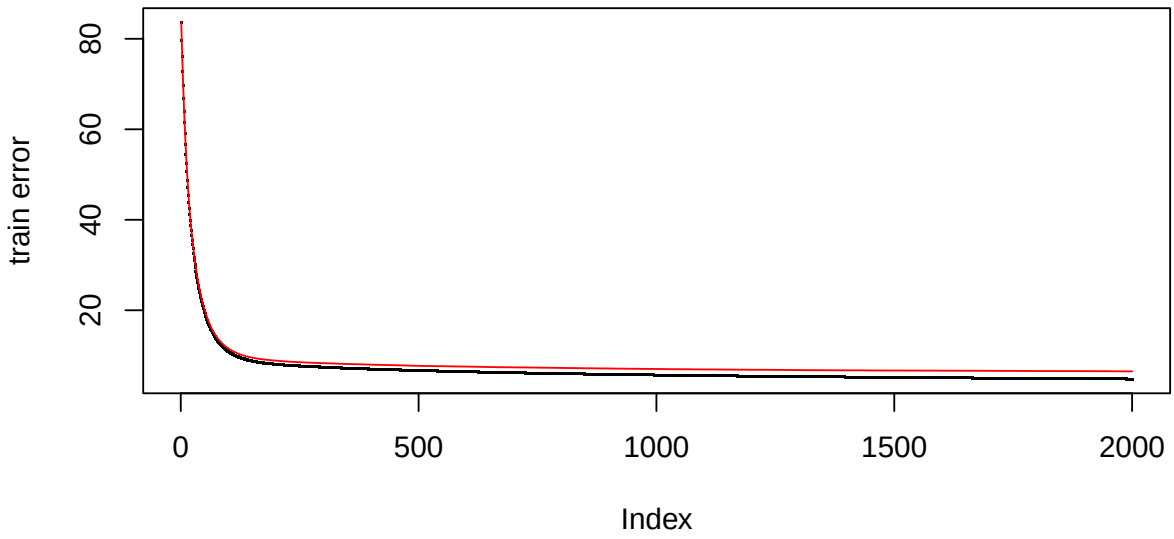
Plot of Training and Testing MSE We also calculate each model's training MSE, cross validation error, and testing MSE for comparison to pick our final boosting model.



According to the plot, we could find that as the shrinkage increases, the training MSE decreases. The reason for that may be we fit the training set better and better when the shrinkage increases, which may lead to an over-fitting. Hence, we need to pick the optimal value of shrinkage parameter by their cross validation error. For 1000 and 2000 trees models, the validation error is gradually decrease when learning rate increase. However, The validation error increases when we have learning rate over 0.45 in 3000 trees model, which means higher shrinkage may not reduce validation error. Additionally, higher learning rate also result in high risk of over-fitting. Hence, pick learning rate around 0.04 would be optimal.



According to our three plots of testing MSE for each number of trees with different shrinkage parameter, we can see that the testing MSE increases or become flatten after x = 0.04, which means we should pick it as our value of shrinkage parameter. For number of trees, '1000' has the largest MSE around 7.2. However, for '2000' and '3000, their MSE are similar, both are around 6.75. To reduce our model complexity and save efficiency, we should pick 2000 as our number of trees. Hence, our final learning rate is 0.04 and 2000 tress will be fitted and we calculate its test MSE for further model comparison.



According to the plot we have, the deviation between validation error and train error is become smaller as we have more iterations.

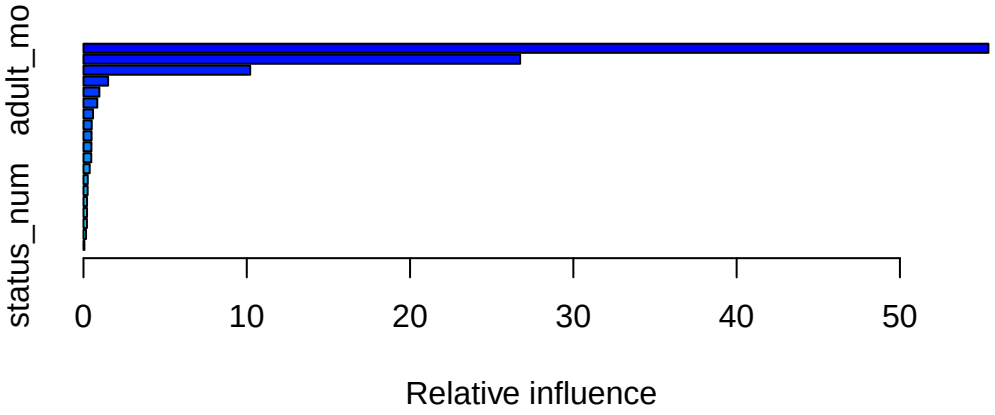


Table 4: Reletive influence for each variable in Gradient Boosting Model

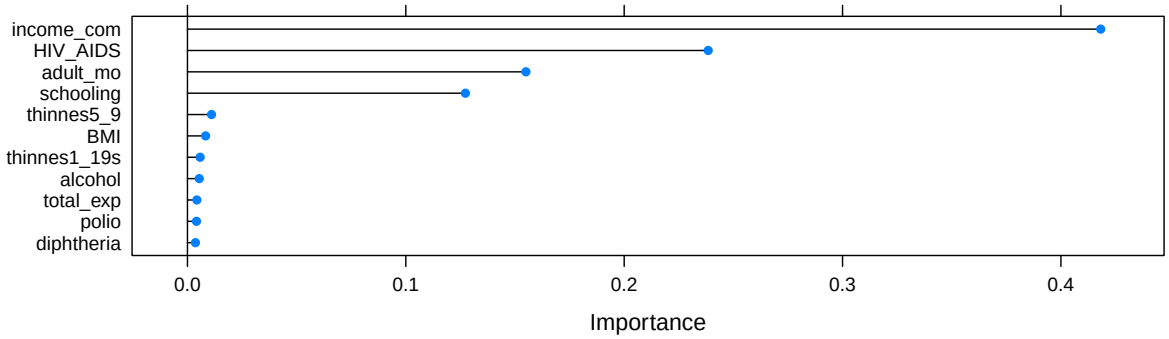
	var	rel.inf
income_com	income_com	55.4197399
HIV_AIDS	HIV_AIDS	26.7417484
adult_mo	adult_mo	10.2092989
total_exp	total_exp	1.5076932
thinnes5_9	thinnes5_9	0.9737532
diphtheria	diphtheria	0.8465734
under_f_dea	under_f_dea	0.5846914
infant_dea	infant_dea	0.5043532
thinnes1_19s	thinnes1_19s	0.4945318
schooling	schooling	0.4853761
alcohol	alcohol	0.4768658
Measles	Measles	0.3812434
GDP	GDP	0.2606405
population	population	0.2545087
polio	polio	0.2190863
BMI	BMI	0.2177754
hepatitis_B	hepatitis_B	0.2099296
percent_exp	percent_exp	0.1573048
status_num	status_num	0.0548861

According to the variable importance plot we have, there is a clear difference in relative influence between variables. There are three most important variables: HDI(income_com), 'HIV/AIDS death', and adult mortality, that dominate our boosting model, which means simpler model may have similar performance. By the table for each variable and their corresponding relative influence in Gradient Boosting Model. The variable that is the most influential is 'income_com', indicating HDI, with relative influence 55.4197399. Variable 'total_exp', indicating health expenditure, has relative influence 1.5076932.

Perform Extreme Gradient Boosting

Based on the wrangled data, we perform extreme gradient boosting model to predict life expectancy. We set up a tuning grid that can help us to perform grid search on eta, max_depth, and nrounds. Based on our data, and 'xgbTree' method, we train our xgb model on the tune grid.

After training, we have our variable importance plot.



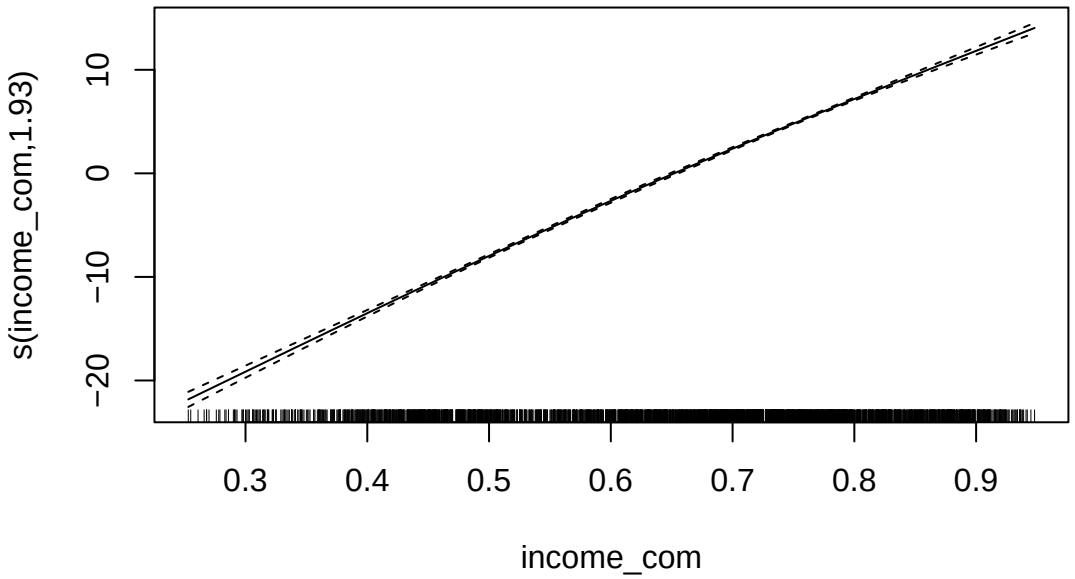
According to the plot, we can see that the difference of importance between variables are pretty clear. Variable 'income_com' is also the most important feature in extreme gradient boosting model. Total expenditure is the 11th important variable. We also find HIV/AIDS deaths and adult mortality, and schooling played an significant role in predicting life expectancy.

Result Section

Comparing all spline models

Table 5: Comparing all R squared of all spline models

models	R_square
total expenditure as smooth terms	0.0379077
income composition of resources as smooth terms	0.7903460
total expenditure as smooth terms adjusted by status	0.2422911
income composition of resources as smooth terms adjusted by status	0.7903013

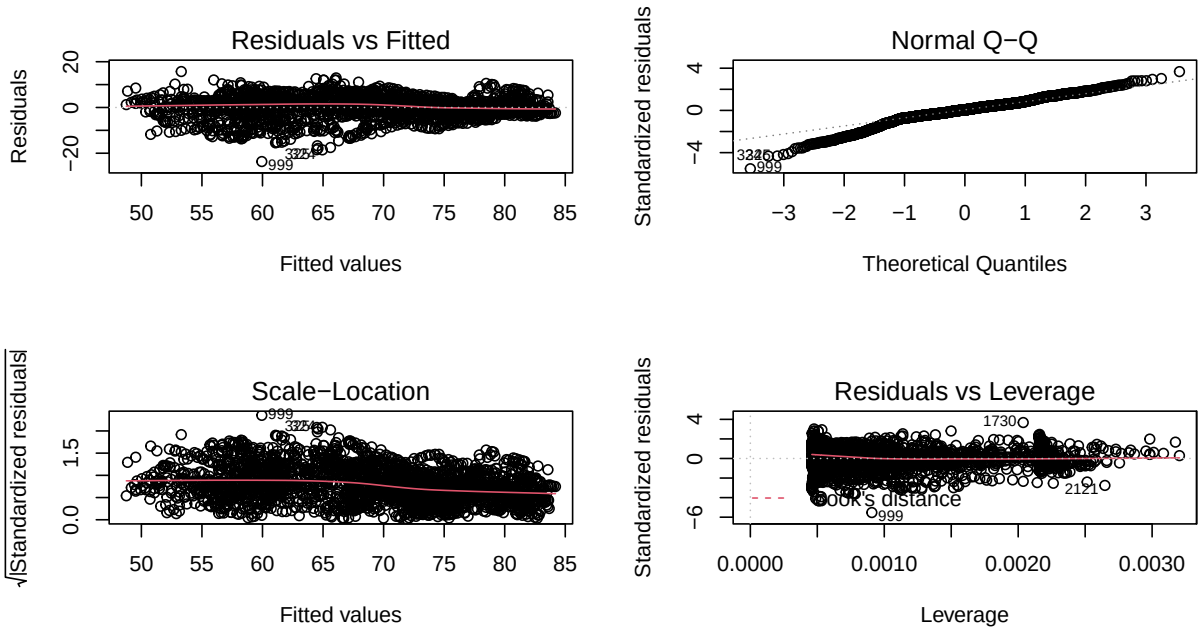


The model of income composition of resources as smooth terms (M6) has the highest R squared value, which is 0.790346. Looking at the spline model we have, the trend is not curvy, which indicate that a liner model may be preferred to reduce overfitting.

Comparing all linear models

Table 6: Comparing all R squared of all linear models

models	R_square
total expenditure as predictor	0.0326109
income composition of resources as predictor	0.7892844
total expenditure and status as predictor	0.2336129
income composition of resources and status as predictor	0.7894635



According to the all R squared value we have for all linear models, the one with income composition of resources and status as predictor (M8) have the highest R squared. Residual vs fitted plot, QQ plot, scale-location plot, and leverage plot were checked for linear model assumptions. Only the QQ plot shown a deviation exist on the left tail, which means normality is slightly violated. Other assumptions are satisfied. Hence, we do have a good fit since assumptions are mostly satisfied.

Comparing all linear mixed models

To compare all of the linear mixed models we have, we need to compare them by likelihood ratio test (lrtest()). We compare two model with different complexity by the p-value we have in the test. If the p-value is smaller than our significant level 0.05, we are able to reject null hypothesis that simpler model have similar prediction accuracy as more complex model, which means picking complex model is more statistically significant. Hence, by our model construction, the complex model in each pair of test is the model that include developing status variable. Then, we compare (M3, M9) and (M4, M10).

Table 7: Linear mixed model comparisons

lrtest	P_value
Likelihhod ratio test between M3 and M9 (Total Expenditure)	0.000000
Likelihhod ratio test between M4 and M10 (HDI)	0.238603

According to the p-value we have, only adding variable status to total expenditure model has a significant improvement. Hence, M9 and M4 would be compared by AIC with final linear and spline models. They cannot be compared by likelihood ratio test again since they does not sharing same predictors.

Comparing picked linear model, picked spline model, and picked linear mixed models

Table 8: Comparing R square for picked linear model and picked spline model

models	R_square
linear model	0.7894635
spline model	0.7903460

Table 9: Comparing AIC for all Regression model

Statistics	AIC.value
AIC of picked linear model M8:	15172.54
AIC of picked spline model M6:	15159.38
AIC of picked linear mixed model: M4	15177.39
AIC of picked linear mixed model: M9	18542.02

According to the table we have, the R squared value for both models are pretty close. Though spline model yields better fit based on the score, a linear model may be better choice since the spline model we plotted is very close to a linear line. Choosing a linear model with almost the same wellness of fitting could reduce over-fitting.

By comparing AIC in Table, though spline model also has the smallest AIC value, most of them have very close AIC. Both our linear mixed model have higher AIC values. Since the linear model is our next-best model and we would like to reduce over-fitting, the linear model with Income composition of resources and status as predictor as predictors is our best model in this section, which means income composition of resources (HDI) has stronger relationship with life expectancy than health expenditure.

Comparing machine learning models by MSE

Table 10: Comparing MSE of all models

models	MSE
Pruned Regression Tree	6.978559
Bagging	3.682146
Random Forest	3.420492
Gradient Boosting	6.890100
Extreme Gradient Boosting	3.557269

According to the MSE table, we can see that Extreme Gradient Boosting model has the smallest test MSE, which is 3.345888. Pruned Regression Tree has the largest MSE 6.978559, indicating a worse fit. Low test MSE shows high performance and low over-fitting. Hence, we may pick extreme gradient boosting model as our final model to predict life expectancy.

Conclusion and Summary

Answering research question

1. Is government health expenditure have greater impact on life expectancy than Human Development Index?

According to the data exploratory plots we have, the relationship between health expenditure and life expectancy is not strong. The models that only contains total expenditure and status highest AIC, which means they fitted badly. However, in most of our plots, the relationship between HDI and life expectancy is strong. We also have pretty well fitted models with HDI as predictor have adjusted R squared over 0.79. Furthermore, in our variable importance plots in machine learning models, HDI is the one of the most important features across all models. Hence, we concluded that HDI have greater impact on life expectancy than the other.

2. Does life expectancy also depends on the development status of the country?

Though including it did improve model performance, adding development status into our model does not have any significant effect according to the model comparison results.

3. How to accurately predict life expectancy by social factors?

Predicting life expectancy by extreme gradient boosting model had the best performance. Variable 'HIV_AIDS', 'income_com', 'adult_mo', and 'schooling' are the most important features to predict life expectancy.

Discussion and limitation

According to the result we have, if the governments aim to increase life expectancy of the population, they should focus on factors that will increase the HDI. Not just spending too much money health development. In may cases, better health treatment and medical technologies does not benefit everyone. For most of the population, even poorer people, their health conditions need much longer time to respond to larger health expenditure than richer people. Additionally, people who would like to search for some counties to stay and try to live longer could choose countries with high HDI, rather than high health expenditure. For government or social organizations that would like to predict local life expectancy, they need to focus on HIV infection records, local HDI, adult mortality, and number of years of schooling. Their values are statistically significant on predicting life expectancy. Fitting a extreme gradient boosting model would yield more accurate values a based on our research result.

Limitations:

1. Since we impute NAs by mean value, we may result in biased standard error, variance, and sample mean. Our estimate may be pulled by other observations.
2. Since we have 16 years of observations per country and we investigate the data-set as a whole, our observations are not totally independent. Also, we ignored the structural difference between countries, like race and climate.
3. Though in model comparison, some spline models have better performance than the linear model, picking a spline model also increase our risk on over-fitting.
4. We only fitted spline models with 3 knots. We should also vary it to compare spline models with different knot numbers.

5. In many of our machine learning models, we can see there are only a few variables that dominate our model, which means, in further study, we could reduce the model complexity but also have similar performance.

Reference

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.