The Most Important Factors in Determining Human Development

Emma Clift & Cathal Mee

11/08/2024

Introduction

We will analyze datasets from the United Nations Development Program, called Human Development Reports, which were collected through the UN's Human Development Data API. These reports include variables of core life measurements that are critical to a country's human development index. These variables ranging from topics in national income, inequality indexes, measures differing by male and females, and environmental factors.

Our Question

The main question we are investigating is: **Which quality of life measurements are most critical to improving a country's human development index?**

Data Wrangling

To answer this question, we wanted to look at all the possible quality of life measurements. Therefore, our first step in data wrangling was to combine every report into one large dataset. The result had a row for each country and a column for each measurement taken in each year, such as hdi_1990, hdi_1991, etc, which totaled over 1000 columns. We pivoted the table to add a `year` column and reduce the dataset to only have one column for each variable. We only wanted to look at data collected between 2016 and 2022, so we eliminated observations taken before 2016. Our final dataset had 6 rows for each country (~1440 rows in total, each showing a different year) and 31 columns to show the recorded measurement for each variable.

Our Dataset

The columns in our data set shows the values of our variables, such as `HDI` (Human Development Index), `LE` (Life Expectancy), `EYS` (Expected Years of Schooling), `GII` (Gender Inequality Index), `MMR` (Maternal Mortality Ratio), and more. Table 1 below includes a summary of our variables including the variable name, type of variable, and units.

According to Table 1, all of our explanatory variables are numerical except year and country which both have multiple levels. Most of our variables are measured by either `percent` or `value`. Some of our variables are different index measures, which are measured by `value`. There are also multiple variables that have a `female` and `male` version.

Table 1: All of the variables available in our dataset, including their variable name, what their meaning, type of variable, explanatory/response, and the units.

Variable	Meaning	Type	Information	Units
HDI	Human Development Index	Numerical	Response variable	Value
Country	NA	Categorical	Explanatory	NA
Year	NA	Categorical	Explanatory	NA
LE	Life Expectancy	Numerical	Explanatory	Years
EYS	Estimated Years of Schooling	Numerical	Explanatory	Years
MYS	Mean Years of Schooling	Numerical	Explanatory	Years
GNIPC	Gross National Income Per Capita	Numerical	Explanatory	Purchasing Power Parity
GDI	Gender Development Index	Numerical	Explanatory	Value
IHDI	Inequality-adjusted HDI	Numerical	Explanatory	Value
Loss	Loss in HDI Value Due to Inequality	Numerical	Explanatory	Percent
GII	Gender Inequality Index	Numerical	Explanatory	Value
MMR	Maternal Mortality Ratio	Numerical	Explanatory	Deaths per 100,000 live
ABR	Adolescent Birth Rate	Numerical	Explanatory	Births per 1,000 women
PHDI	Planetary Pressures-Adjusted HDI	Numerical	Explanatory	Value
HDIF	Human Development Index Female	Numerical	Explanatory	Value
LEF	Life Expectancy Female	Numerical	Explanatory	Value
EYSF	Expected Years of Schooling Female	Numerical	Explanatory	Value
INEQEDU	Inequalty in Education	Numerical	Explanatory	Percent
INEQINC	Inequality in Income	Numerical	Explanatory	Percent
CO2	Carbon Dioxide Emissions Per Capita	Numerical	Explanatory	Tonnes
MF	Material Footprint Per Capita	Numerical	Explanatory	Tonnes
POPTOTAL	Population Total	Numerical	Explanatory	Millions
HDIM	Human Development Index Male	Numerical	Explanatory	Value
LEM	Life Expectancy Male	Numerical	Explanatory	Value
EYSM	Expected Years of Schooling Male	Numerical	Explanatory	Years
SEF	Some secondary Education Female	Numerical	Explanatory	Percent
SEM	Some Secondary Education Male	Numerical	Explanatory	Percent
PRF	Share of seats in Parliament Female	Numerical	Explanatory	Percent
PRM	Share of Seats in Parliament Male	Numerical	Explanatory	Percent
LFPRF	Labor Force Participation Rate Female	Numerical	Explanatory	Percent
LFPRM	Labour Force Participation Rate Male	Numerical	Explanatory	Percent
INEQLE	Inequality in Life Expectancy	Numerical	Explanatory	Value

Statistical Summary

In Table 2, we have the minimum, mean, and maximum values for the variables that are not adjusted HDI measures or specific to male or female. From Table 2, we see that the spread of variable values ranges from just above 0 to over 80 thousand. Since most variables have different units, we can't compare the values of `GNIPC` (Gross National Income Per Capita) to `ABR` (Adolescent Birth Rate) without accounting for the variable ranges. According to the Table, we see that most variables are either normally distributed. `ABR` (Adolescent Birth Rate), `CO2` (Carbon Dioxide Emission Per Capita), `MF` (Material Footprint Per Capita), and `GNIPC` (Gross National Income Per Capita) all are skewed to the right.

Table 2: The minimum, mean, and maximum values from variables that are not adjusted HDI index measures or specific to male or female.

Variable	Meaning	Minimum	Mean	Maximum
HDI	Human Development Index	0.372	0.7281	0.967
LE	Life Expectancy	52.040	72.4100	84.820
EYS	Expected Years of Schooling	6.135	13.6230	22.856
MYS	Mean Years of Schooling	1.246	8.9240	14.256
GNIPC	Gross National Income Per Capita	712.000	20341.0000	88761.000
Loss	Loss in HDI Value Due to Inequality	4.752	20.5250	44.020
MMR	Maternal Mortality Ratio	1.664	138.6730	1135.414
ABR	Adolescent Birth Rate	1.814	46.4260	179.765
CO2	Carbon Dioxide Emissions Per Capita	0.031	4.1800	25.376
MF	Material Footprint Per Capita	1.181	12.9300	49.563

Table 3: The mean values for variables with a male and female measure.

Variable	Meaning	Female.Mean	Male.Mean
HDI	Human Development Index	0.707	0.742
LE	Life Expectancy	75.070	69.800
EYS	Expected Years of Schooling	13.834	13.441
SE	Population with Secondary Education	60.193	65.605
PR	Share of seats in Parliament	24.670	75.330
LFPR	Labor Force Participation Rate	50.700	70.900

When we look at Table 3 above, we can compare the mean values of female and male measures. According to the table, the Human Development Index for female's is slightly lower than males by 0.035. Females have a life expectancy around 6 years more than male. The `EYS` (Expected Years of Schooling) for males and females is similar, but the female measure for `SE` (Population With Secondary Education) is around 5.5% lower. The female measure for `PRF/PRM` (Share of Seats in Parliament) are around 50% lower than the males and around 20% lower for `LFPRF/LFPRM` (Labor Force Participate Rate).

Exploratory Data Analysis

For our exploratory data analysis, we chose to focus on what variables are the most important and show the most interesting information with our response variable, `Human Development Index`. Starting with a histogram, we investigating the spread of the `Human Development Index` from our dataset. Since the `HDI` (Human Development Index) does not follow a normal distribution, when we further interpret our variables and make our models, we need to take into account the distribution of `HDI`.

The mean human development index is 0.728. Figure 1a is skewed to the left with there being peaks in the frequency of countries' `HDI` at around 0.75 and 0.95. This histogram is important in understanding our dataset's skew in `HDI` as the spread of `HDI` is not normally distributed. Figure 1b showcases how the female ad male distributions of life expectancy are different, as female life expectancy has a higher mean than the male measure.

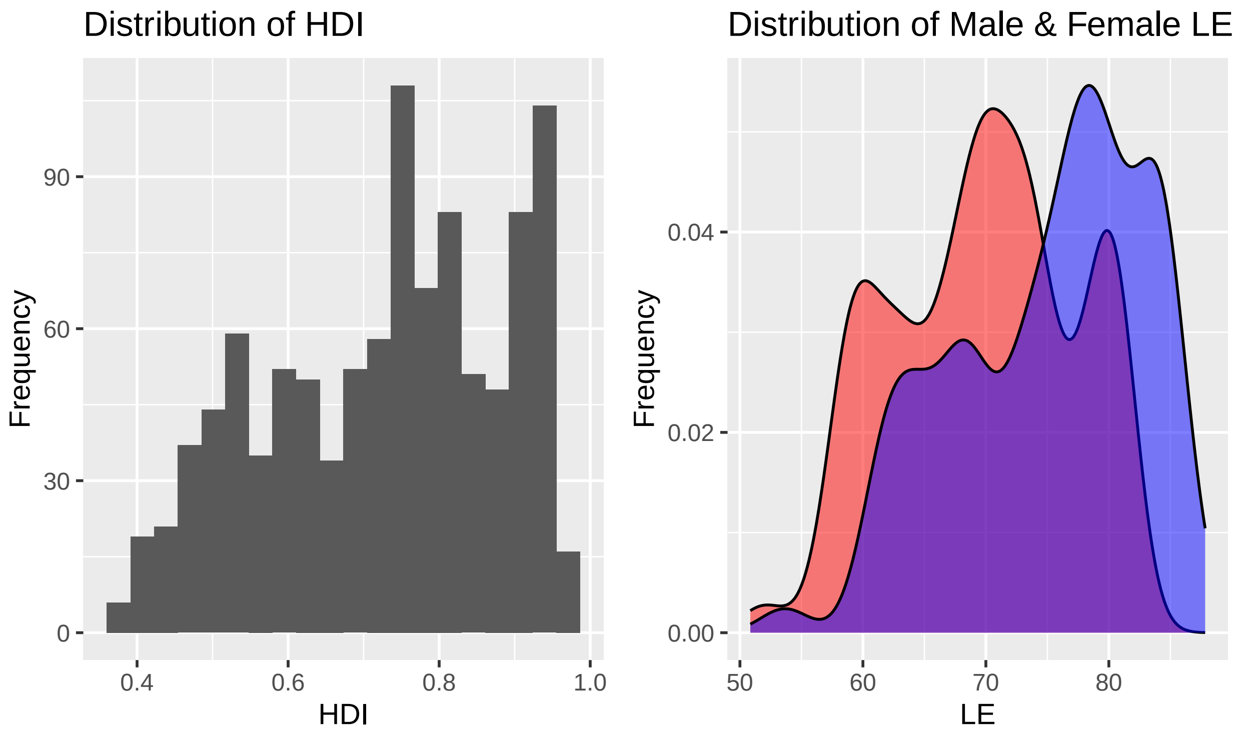

Figure 1: a) Histogram showing the distribution of all countries' HDI throughout the years 2016 to 2022 with bins = 20 b) Density plot where the red curve represents the male's life expectancy distribution and the blue curve represents the female's life expectancy distribution.

As shown in Figure 2.a, there is a logarithmic relationship between the `GNIPC` (Gross National Income Per Capita) and the `HDI` Human Development Index. As `GNIPC` increases, `HDI` is also expected to increase, but as the `GNIPC` increases above 25,000 PPP, the increase in `HDI` is slower than before. This relationship is important because `GNIPC` plays a significant role in determining a country's `HDI`, as seen by the strong positive logarithmic relationship between the two. One important thing to consider is that a country can have a high `GNIPC`, but not necessarily have high measures of other factors, so we need to create a more complex picture beyond `GNIPC`.

As shown in Figure 2.b, there is a positive linear relationship between life expectancy and human development index. As the life expectancy increases, the human development index also increases. This is important because it could be a major predictor of a country's `HDI` (Human Development Index).

As shown in Figure 2.c, there is a positive, linear relationship between the two variables. As `EYS` (Expected Years of Schooling) increases, the `HDI` is increases as well. This makes sense the `EYS` is directly correlated with the education access in particular countries. If a country has a higher access and standard to education, their `HDI` will likely be higher. This is because the country's educational system has a large impact on their `HDI`.

In Figure 2.d, there is a strong negative linear relationship between `ABR` (Adolescent Birth Rate) and `HDI`. This is important because it could signify that `ABR` could be a strong predictor for `HDI`.

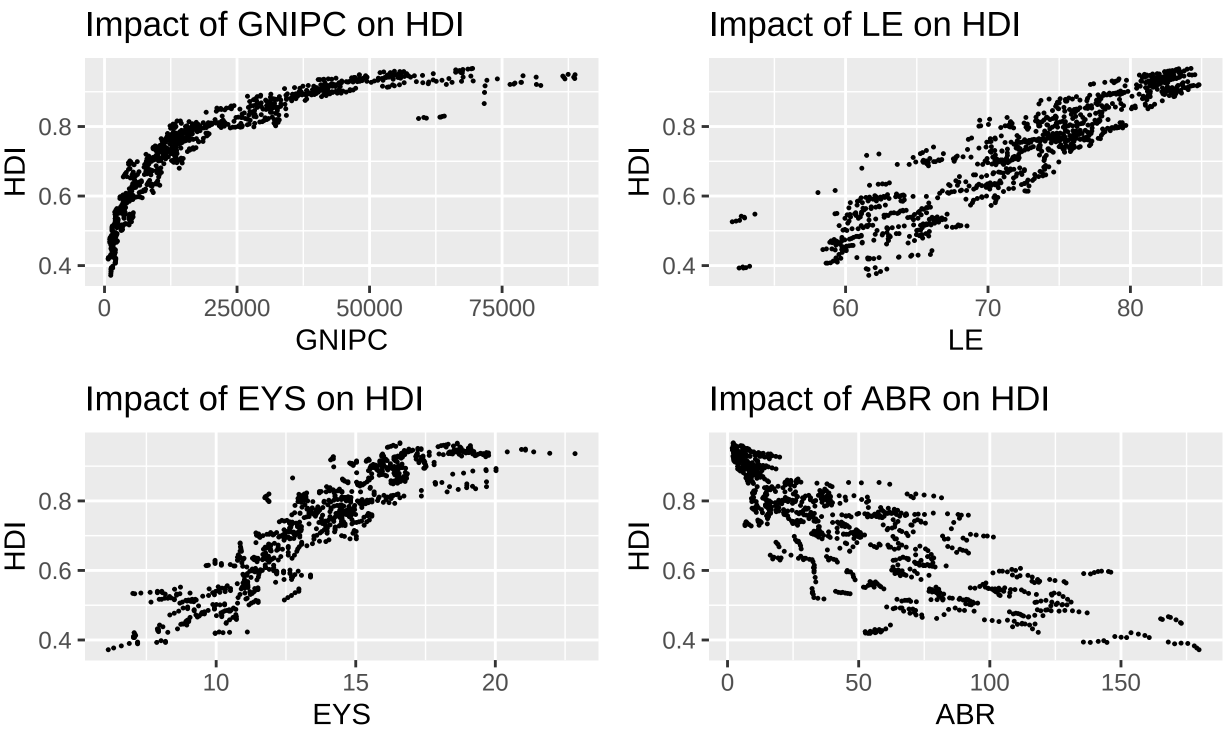

Figure 2: a) A higher Gross National Income Per Capita is associated with a higher Human Development Index. Data are based on yearly measurements taken between 2016 and 2022 for over 190 different countries. GNIPC is measured in Purchasing Power Parity, which is a currency conversion rate used to compare the amount of money necessary to buy the same goods in different countries. The HDI is measured on a scale from 0-1. b) A higher Life Expectancy is associated with a higher Human Development Index. Data are based on yearly measurements taken between 2016 and 2022 for over 190 different countries. Life Expectancy is measured in years, and Human Development Index is measured on a scale from 0 to 1, where 1 indicates higher development. c) A higher Expected Years of Schooling is associated with a higher Human Development Index. Data are based on yearly measurements taken between 2016 and 2022 for over 190 different countries. Expected years of Schooling is measured in years, and Human Development Index is measured on a scale from 0 to 1, where 1 indicates higher development. d) A higher Adolescent Birth Rate (ABR) is associated with a lower Human Development Index. Data are based on yearly measurements taken between 2016 and 2022 for over 190 different countries. ABR is measured in births per 1,000 women, and ranges from 0 to around 175. The HDI is measured on a scale from 0-1.

Model Building

Lasso

The first method we chose to answer our research question is a LASSO model. The main reason for this was because we have a large number of explanatory variables. This is a regression model to predict the value of `HDI` based on ALL of the explanatory variables. We used a 10-fold cross validation dataset and grid to optimize the penalty value for our LASSO model.

```r
# Create the training and testing dataset
set.seed(12345)
hdi_split <- initial_split(hdi_tidy, prop=0.8)
hdi_train_tbl <- training(hdi_split) |>
  select(!country & !ihdi & !hdif & !hdim & !phdi)
hdi_test_tbl <- testing(hdi_split) |>
  select(!country & !ihdi & !hdif & !hdim & !phdi)

#Define the lasso model
lasso_model <-
  linear_reg(mixture = 1, penalty=tune()) |>
  set_mode("regression") |>
  set_engine("glmnet")

#Define the lasso recipe
lasso_recipe <-
  recipe(formula = hdi ~ ., data = hdi_train_tbl) |>
  step_dummy(all_nominal_predictors()) |>
  step_zv(all_predictors()) |>
  step_normalize(all_predictors())

#Define the lasso workflow
lasso_wf <- workflow() |>
  add_recipe(lasso_recipe) |>
  add_model(lasso_model)

#Fit the basic model to our data (this assumes a penalty of 1)
lasso_fit <- fit(lasso_wf, data = hdi_train_tbl)

#Create the graph of our variable coefficients
#This allows us to determine the range for our penalty grid
lasso_var_graph <- lasso_fit |>
  extract_fit_engine() |>
  autoplot()
```

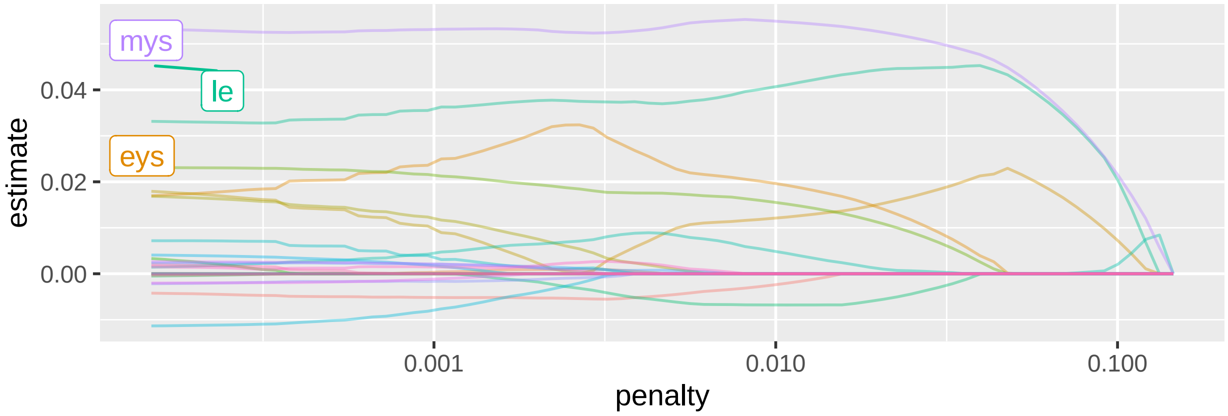

Figure 3: Plot representing the relationship between the coefficient's estimate on a range from around -0.02 to 0.06 and penalty on a logarithmic scale from 0 to around 0.1.

When looking at Figure 3, the majority of variables do not have a very large estimate and go towards 0 before a penalty of 0.010. There are still 6 variables that do not follow a similar pattern and either have a high estimate before going down to 0, or follow a non-linear/logarithmic pattern. For example, the variable `EYS` (Expected Years of Schooling) starts with an estimate of 0.02 and it rises to over 0.03 before eventually going to 0. The two highest variables regarding their estimate are the `MYS` (Mean Years of Schooling) and `LE` (Life Expectancy). Using this information, we created 10 folds for our cross validation, and choose a penalty grid from penalty(c(-5,1)) with 10 levels.

```r
# Create the folds and penalty grid for our cross validation
lasso_fold <- vfold_cv(hdi_train_tbl, v = 10)

# We chose these values for the penalty range due to the graph above.
penalty_grid <-
  grid_regular(penalty(c(-5,-1)), levels = 10)

# Tune the penalty grid
tune_lasso_res <- tune_grid(
  lasso_wf,
  resamples = lasso_fold,
  grid = penalty_grid
)
```

After tuning our model with the penalty grid, we looked at what penalty optimized our R-squared and RMSE values. According to Figure 4, the penalty value that has the lowest RMSE value is around 0.00001 with around the highest R-squared value.

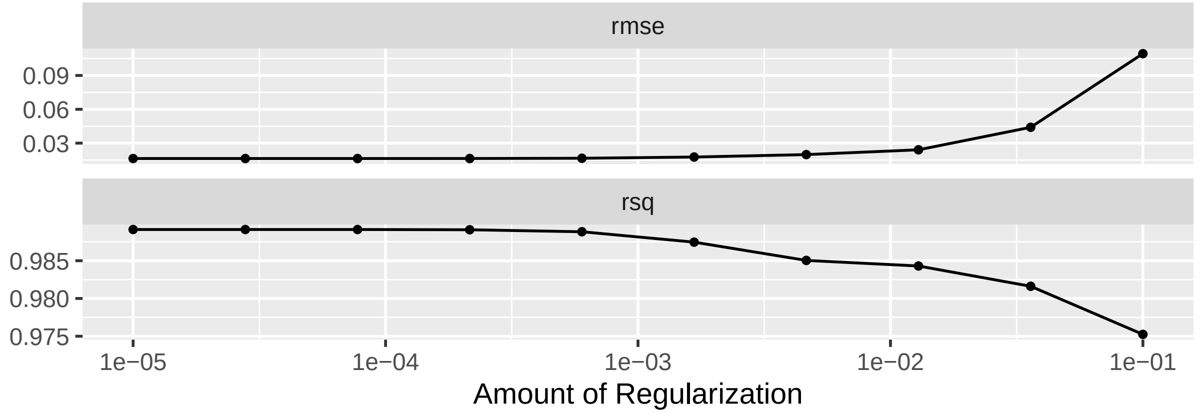

Figure 4: RMSE and R-squared values vs the amount of regularization ranging from around 1e-05 to 1e-01. RMSE values range from 0.00 to 0.11 and R-squared values range from 0.973 to 0.994.

We selected this best penalty and finalized our workflow, fitting the final model using resampling.

```r
# Select our best penalty
best_lasso_penalty <- select_best(tune_lasso_res, metric = "rmse")
#Finalize the workflow with our optimal penalty and fit the model
lasso_final_wf <- finalize_workflow(lasso_wf, best_lasso_penalty)
#Fit the lasso model workflow using resampling
lasso_final_fit <- fit_resamples(lasso_final_wf, lasso_fold)
metrics <- collect_metrics(lasso_final_fit)
```

After fitting our lasso model with our optimized penalty value and the 10-fold cross-validation dataset, our $RMSE = 0.016$ and $RSQ = 0.989$. The 5 variables with the highest variable importance in this model are `MYS` (Mean Years of Schooling), `LE` (Life Expectancy), `GNIPC` (Gross National Income Per Capita), `EYSM` (Expected Years of Schooling Male), and `EYS` (Expected Years of Schooling).

Random Forest Model

The second model we decided to use is a Random Forest Model. Similar to the lasso model, this is a regression model to predict the value of HDI based on ALL of the explanatory variables. We used a 10-fold cross validation dataset and grid to optimize the `mtry` parameter (number of variables randomly sampled at each split).

After tuning our model with the grid of mtry values (ranging 1 to 25, since we have 25 explanatory variables), we looked at what values optimized our RMSE. As shown in Figure 5 below, the # Randomly Selected Predictors (mtry) that gives us the best RMSE value is 11 where our RMSE is 0.006 and `tree_depth` is 10.

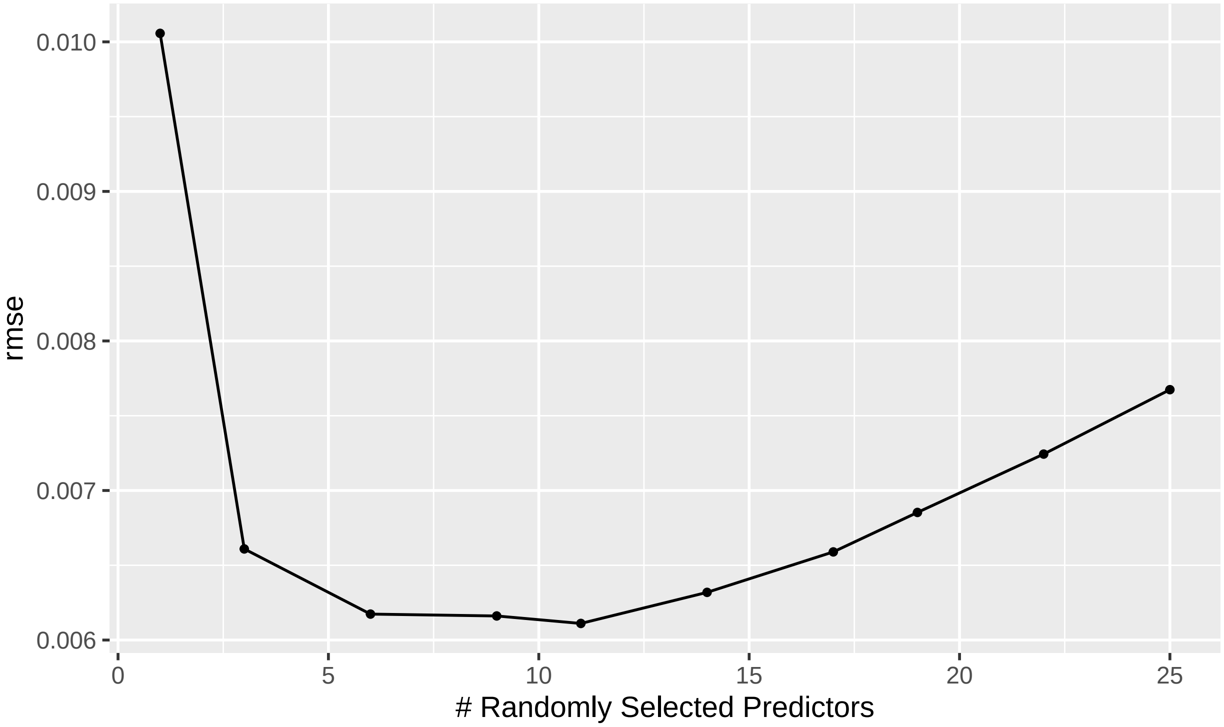

Figure 5: RMSE values vs number of selected predictors ranging from 1 to 25. RMSE values range from 0.00 to 0.10.

In table 4, we have all of the mtry values with their RMSE values and standard error. We choose `mtry` of 11 as the most optimal value through the one-standard-error rule to achieve the simpliest Random Forest model that still optimized our RMSE.

Table 4: This table shows the mean RMSE for all mtry values within the range of our grid.

mtry	.metric	.estimator	mean	n	std_err	.config
11	rmse	standard	0.0061110	10	0.0003090	Preprocessor1_Model05
9	rmse	standard	0.0061610	10	0.0003935	Preprocessor1_Model04
6	rmse	standard	0.0061735	10	0.0002544	Preprocessor1_Model03
14	rmse	standard	0.0063189	10	0.0003438	Preprocessor1_Model06
17	rmse	standard	0.0065894	10	0.0002779	Preprocessor1_Model07

Table 5: The Five most important variables in the Random Forest Model based on Variable Importance of Projection.

Variable	Name	VIP
GNIPC	Gross National Income Per Capita	11.80
INEQLE	Inequality in Life Expectancy	1.61
MYS	Mean Years of Schooling	1.53
LEF	Female Life Expectacy	1.39
EYSF	Female Estimated Years of Schooling	1.27

After fitting our Random Forest model with our optimized `mtry` value and the 10-fold cross-validation dataset, our $RMSE = 0.007$ and $RSQ = 0.998$. The 5 variables with the highest variable importance in this model are `GNIPC` (Gross National Income Per Capita), `INEQLE` (Inequality in Life Expectancy), `MYS` (Mean Years of Schooling), `LEF` (Female Life Expectancy), `EYSF` (Expected Years of Schooling Female).

Model Refinement

First, we ran our models based on all variables in the dataset. Our models determined that the most important variables for predicting `HDI` were `HDIF`, `HDIM`, `IHDI`, and `PHDI`, which are the `HDI` measurements for specifically males/females, inequality adjusted `HDI`, and Planetary Adjusted `HDI.` These variables are either subsets of the overall `HDI` or an adjusted `HDI` index. Due to this, we decided to eliminate these variables from our testing/training datasets, since our models were able to predict `HDI` using the similar measures. Those results don't answer our question, since they are subsets of `HDI.` We are the most focused on how characteristics of a country, such as life expectancy or estimated years of schooling, impacts a country's human development index. For both our LASSO and Random Forest models, we didn't use these variables so we were able to see the effect of the variables of most interest to us. We did not filter these variables when wrangling our data. To remove these variables from our model, we did not select variables `HDI`, `HDIF`, `HDIM`, `IHDI`, `PHDI`, and `country` from our testing and training datasets used for our models. Out of all of the variables in our Random Forest model, there are some relationships between the variables, but none of them are completely correlated with each other. For example, mean years of schooling and expected years of schooling are related, but their values are independent of each other. Due to this, both variables in relationships similar to that are included in our model. Our five most important variables in our Random Forest model are shown in table 5 above (table 5 is shown above the Model Refinement header).

Table 6: Fit of the LASSO and Random Forest models by R-squared and RMSE values.

Model	RSQ	RMSE
LASSO	0.99	0.016
Random Forest	0.998	0.007

Conclusion

The resulting RMSE and r-squared values from fitting our models to the 10-fold cross validation dataset are shown in Table 6 below. The r-squared for our LASSO model is 0.99. This means that 99% of the variability in `HDI` can be explained by our explanatory variables. The r-squared for our Random Forest model is 0.998. This means that 99.8% of the variability in `HDI` can be explained by our explanatory variables. The RMSE value for our LASSO model is 0.0164 and for our Random Forest model is 0.007. The Random Forest model has a higher r-squared and lower RMSE value, meaning that this model is a better fit for predicting `HDI` values. In table 6, we have the fit of the LASSO and Random Forest model (Table 6 is seen above in the Conclusion header).

Since we choose to use the Random Forest Model, after fitting the Random Forest model to the testing dataset, the resulting RMSE and r-squared values are the following: RMSE = 0.00625 and R-squared = 0.99829. Additionally, in figure 6 we have the top 5 most important variables of our Random Forest model. Since we choose this model, these are the variables that give us the most insight on what factors are the most important in the Human Development Index.

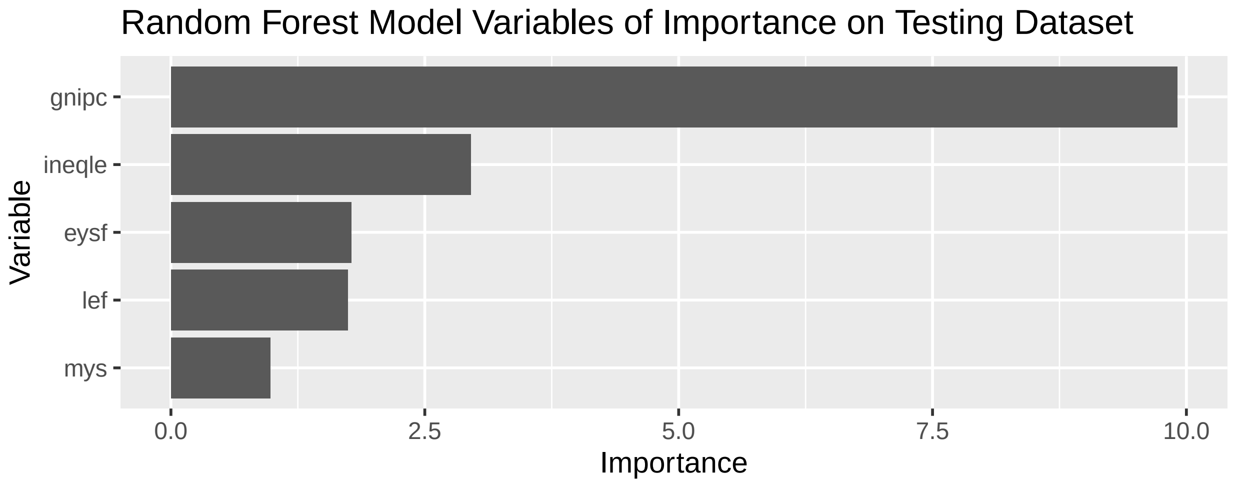

Figure 6: The top 5 variables in our Random Forest model are Gross National Income Per Capita, Inequality of Life Expectancy, Estimated Years of Schooling for Female, Life Expectancy Female, Mean Years of Schooling, and Loss of HDI Due to Inequality

Based on our models, we concluded that the most important measures to improve a country's HDI are Gross National Income Per Capita, Inequality of Life Expectancy, Estimated Years of Schooling for Female, Life Expectancy Female, Mean Years of Schooling, and Loss of HDI Due to Inequality. This makes sense because schooling and `GNIPC` are necessary foundations for many aspects of life. Employment rate and household income might be important statistics, but the root cause lies in schooling, and the root of schooling is the income of a country. Life expectancy is one of the most telling statistics about a variety of core aspects of human society, and a low life expectancy is a strong indicator for the existence of many problems. It makes logical sense that our model picked out these fundamental issues in society as the most influential factors for human development.