Derin Sezercan

ds3789

May 7, 2022

Data Mining Final Project - Determining Indicators that Project Development

Abstract

The initial motivation of this project is to shed light into the long-debated topic of determining indicators that project future development by assessing the effects of currently determine indicators by the World Bank. By the end of this data mining project institutions that collect data on countries and macroeconomic factors like the IMF, IFC, World Bank, UN, OECD, FED, and central banks of individual countries would be highly interested since this project will try to generate an algorithm that classifies countries and provide common ground for comparing countries using various features that determine a country's disposition. This discovery of indicator effects and pattern analysis is considered data mining as I will use various model fittings to find the best algorithm for my conclusion. In economics comparing markets and economies around the world is controversial and there isn't a consensus among economists since they have to use different comparison metrics for different countries and different research topics. This algorithm would be an alternative approach to comparing global economies as it will be able to train a model using various cross-sectional data. The addition of these unconventional metrics for economic comparison will be a new approach in this field. I believe that the algorithms would shed light into the effect of established economic factors and alternative metrics (environmental, demographic, health, education etc.) on macroeconomic development. Additionally, the estimated parameters of the features that will come out from this model would enable central banks to determine important indicators to make decisions on what economic policies to implement and how to change funding preferences within governments to advance their country's economy to become a developed country. The data I used were from, the World Bank Development Indicators data[1] on 36 OECD countries[2] were used. Classification data of development level was from the United Nations country classification file[3]. After doing my work, I realized that limited data is hard to process thus I would try to use a different dataset

[1] https://datacatalog.worldbank.org/search/dataset/0037712

[2] stats.oecd.org and https://www.oecd.org/newsroom/oecd-welcomes-costa-rica-as-its-38th-member.htm#:~:text=The%20OECD's%2038%20members%20are,Norway%2C%20Poland%2C%20Portugal%2C%20Slovak

[3] UNITED NATIONS DEPARTMENT FOR ECONOMIC AND SOCIAL AFFAIRS. World economic situation and prospects 2020. UN, 2020.

without missing values. Usually, such datasets are accessible in return of a fee so I had to utilize public data.

Introduction

Economists have been looking into methods to address and better acknowledge the factors affecting country-based development. The question has been important for macroeconomically involved organizations like the IMF and World Bank while it has been important for central banks of countries to assess areas that need improvement. Being in the developed economies league indicates that a country is good for business and trade so countries aim to improve their economies to be able to have a role among the greater economies which would spur economic activity. The potential value generated from this project can be saved time on endless macroeconomic research that would be put in to determine a country's needs and also reduced uncertainty of economic indicators as the models will be selecting the optimal indicators and their effects.

According to the World Bank[4] in order to assess the development level of a country there are 19 indicator categories that should be evaluated:

1)Agriculture/Rural Development

2)Aid Effectiveness

3)Environmental/Climate Change

4)Economic Growth

5)Education

6)Energy

7)Debt

8)Financial Sector

9)Gender

10)Health

11)Infrastructure/Transportation

12)Poverty/Inequality

13)Private Sector

14)Public Sector

15)Science & Technology

[4] https://data.worldbank.org/indicator

16)Social Development

17)Labor

18)Trade

19)Urban Development

In this project I am going to get features representing these development indicator categories from the World Bank WDI dataset for 36 OECD countries. This WDI dataset is considered a large dataset since it contains about 1400 features/indicators that address the data and needs cleaning and adjusting before data mining can be conducted. For the sake of conciseness and accuracy I am working with OECD countries. Selecting OECD countries will help my models to account for the other exogenous parameters that I am not taking into account like the political systems within the countries and the type of economy (open, close, socialist, capitalist etc.). All OECD countries are typically democratic countries that support free-market economies (although the freedom levels vary, this is the best group I can get with data available for public use) which standardizes important political and economic data for this algorithm. From an Economics - Political Science perspective in order to determine an economy's development level it should be imperative to include politic, international relations, and human rights indicators. Even though the World Bank doesn't emphasize such necessity, Frieden's research paper on the interactions between politics, economics, and other realms[5] acknowledges the strong impact political leadership and decision-making has on the economy. For example, if a country is officially a dictatorship, then the algorithm would have to account for freedom since freedom does have a significant effect on economic decision-making thus economic development as Frieden's research also supports. For this project since all OECD countries are considered to be in similar dispositions politically and strategically there is no need to include such parameters. If this algorithm were to be applied for all countries in the world, then parameters that would account for political dispositions, economic structure of each country, and international relations should be included.

I got the information on OECD countries from stats.oecd.org. Although South Korea and Slovakia are OECD countries the World Bank dataset doesn't have any data for these countries so I will have to use the data available to me but the ideal data set in this case would include all OECD countries or all countries that is of interest if a group is specified.

For people working with this data, it is important to point out the misleading results data snooping could cause. If researcher decides to perform statistical inference or data mining after looking at the data this would yield biased results. Since in this project I am familiar with the country specific economic data I didn't look into individual country data visuals to avoid this bias. I will hard-code the development indicators for countries and then evaluated the datapoints.

For the data mining part of the project, I will compare all models I have fitted with the classification error, precision, recall, and sensitivity metrics of predicted data. Based on the comparison graph I will choose the best model. After selecting the best model I will run the final

[5] Frieden, Jeffry. "The political economy of economic policy: we should pay closer attention to the interactions between politics, economics, and other realms." Finance & Development 57, no. 002 (2020).

model with the most frequently identified significant indicators(which I will get from comparing all my models) and run the model based on only one year which will be 2008(The Great Recession). I am using year first as an explanatory variable to see is the year indicator has a significant effect on the development level prediction, then I will use the data from only the crisis year, 2008, to assess whether my selected best fitting model works better when year and indicators are specified. Specifying year might have a better effect since each year the technology, international relations, and political events differ and affect the OECD countries in similar ways. I am trying to evaluate whether fitting the model per annum would yield better predictions. I think it is also important to determine how the model evaluates data during different periods because when this model is reproduced for country analysis worldwide it would be imperative to account for a crisis yar which would have a negative effect on many indicators and a prosperous year which would pump up the economic indicators. These shifts, when year base data is not accounted for, could be misleading on determining a country's level.

While it is evaluated in the conclusion in detail, my expectations on the year variable weren't accomplished by the models as the year variable was insignificant. As I fitted RandomForest algorithm to the 2008 data and all year data with selected indicators I found out that the model with all year data fits the model better. I had assumed that accounting for year would enable researchers to account for yearly events yet the analysis I have done doesn't provide any evidence to prove this assumption.

Data Cleaning, Wrangling, and Exploration

For this project I am only selecting years between 2000 - 2021 to account for technological and world-wide development since there has been many incidents over history and if I include all years I would also have to account for historical effects. This data will only entail the 21st century and OECD countries since this specific group has a higher chance of fitting the algorithm accurately and extracting information logically. I am first going to use the year variable as a feature to evaluate whether years have a certain effect on a country's development since with each new year there are new inventions, policies, technological developments, diseases, etc. After analyzing year's effect as a feature, I will try to fit the best model on to designated world-wide financial crisis years and designated world-wide prosperous years. I will do this analysis to observe whether my algorithm is better at predicting during crisis years or prosperous years which would provide a better understanding of the machine learning process and would enable my audience to use this algorithm for cases that this algorithm proves to be most accurate.

Also, I will pivot the data table so that all indicators designated by the World Bank and the years which are going to be features will be explanatory variables in my models. All indicators will be columns and year will also be a column.

The data that specifies which OECD countries are developed and developing are from the UN classification document "World Economic Situation and Prospects 2020"[6].

Greece and Turkey are categorized as economies in transition so for this project I am going to assume that it is a developing economy. For this part I am hard coding the explanatory variable of the project. Developing countries=0, developed countries=1.

It is also important to get rid of all columns that have mostly “na” values because those metrics wouldn't provide any insight for us and the artificial data generation with knn means wouldn't suffice. For this “na” cleaning part I will delete the columns that have “na” values more than or equal to half of the total data for each indicator/feature, I am omitting “na” values more than half but while regenerating this project other researchers can use other numbers such as 3/4 or 1/5 depending on the accuracy, they want to get for the knn mean data generation for missing values. In this project I want to have less na values so that knn means could impute missing data points better before fitting my models. So, I dropped all columns that had “na” values more than or equal to half the data for each indicator and I also dropped them from the indicator name data frame that provides name and explanation info for each indicator. Now that all my data is cleaned, I can look into the variables in detail and found that the data are stored as percentages, dollars or scales/indexes.

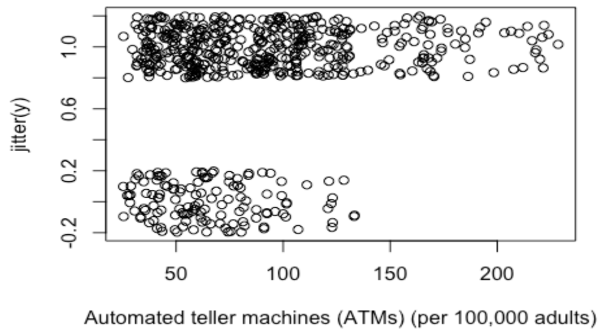
From the data exploration I realized that some data are in LCU which means in local currency, I will use local currency vs USD rate data to convert all LCU data to USD. I will get the exchange rates from the World Bank data in column "PA.NUS.FCRF" which is the "Official exchange rate (LCU per US$, period average)" period average is yearly average in this case and I will fill in the missing exchange rate values from OECD yearly exchange rate data[7] and exclude that column since we are going to account for it with all the data we are converting to USD. 2021 rate was missing for Turkey, so I got the 2021 yearly avg from ExchangeRates.org[8]. I multiplied all LCU data with the corresponding exchange rates to be able to have all data in USD.

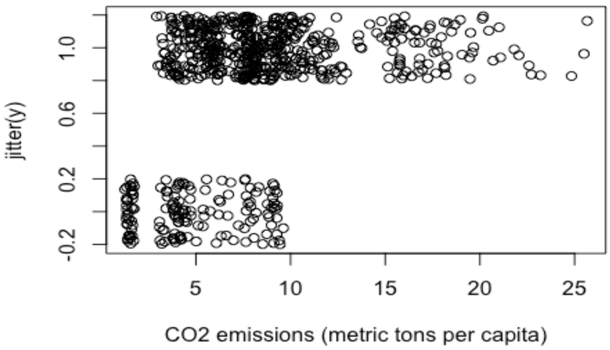
[6] UNITED NATIONS DEPARTMENT FOR ECONOMIC AND SOCIAL AFFAIRS. World economic situation and prospects 2020. UN, 2020.

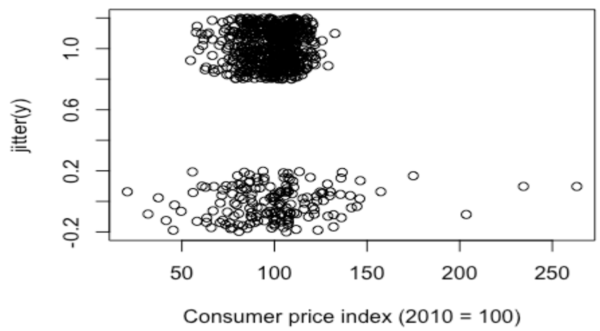
[7] OECD (2022), Exchange rates (indicator). doi: 10.1787/037ed317-en (Accessed on 28 April 2022)

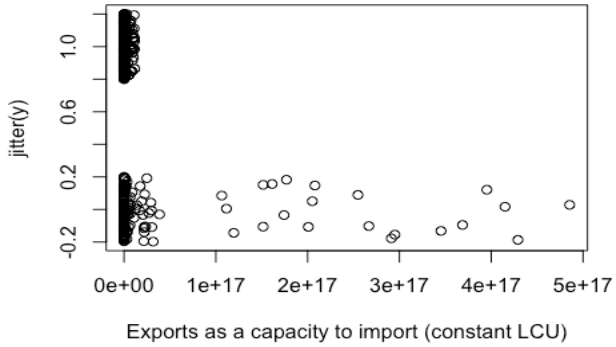
[8] “US Dollar (USD) to Turkish Lira (Try) Exchange Rate History.” Exchange Rates. Accessed May 7, 2022. https://www.exchangerates.org.uk/USD-TRY-exchange-rate-history.html.

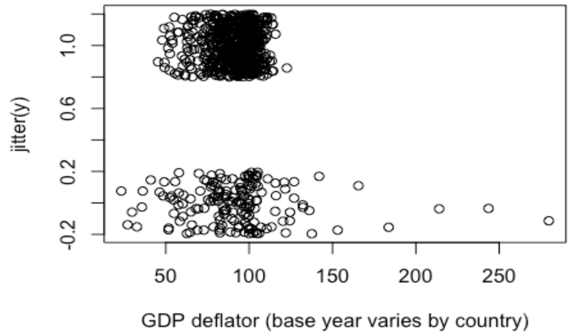
Visualizing Random Data That is Not Collected in % or USD:

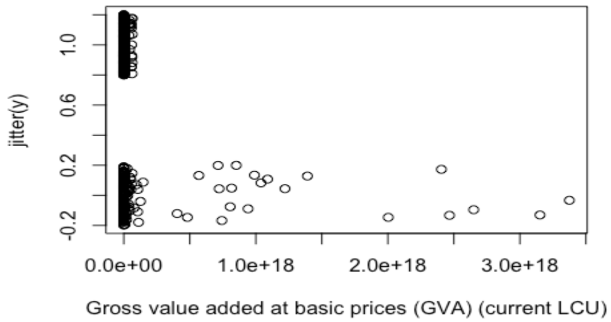


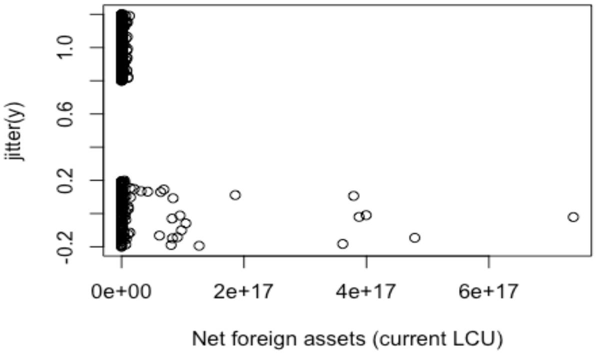


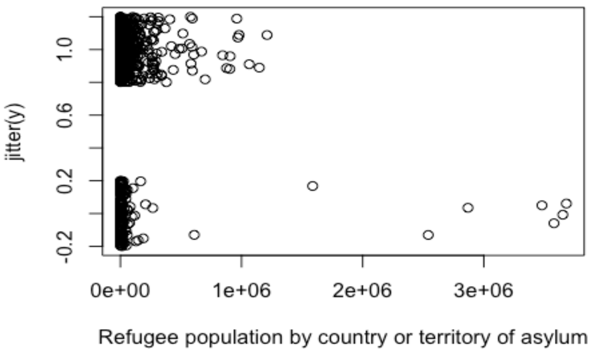


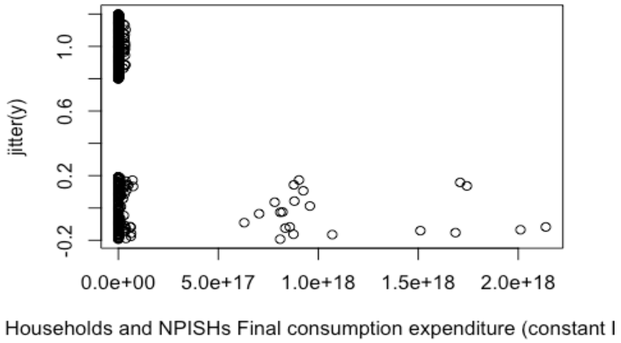






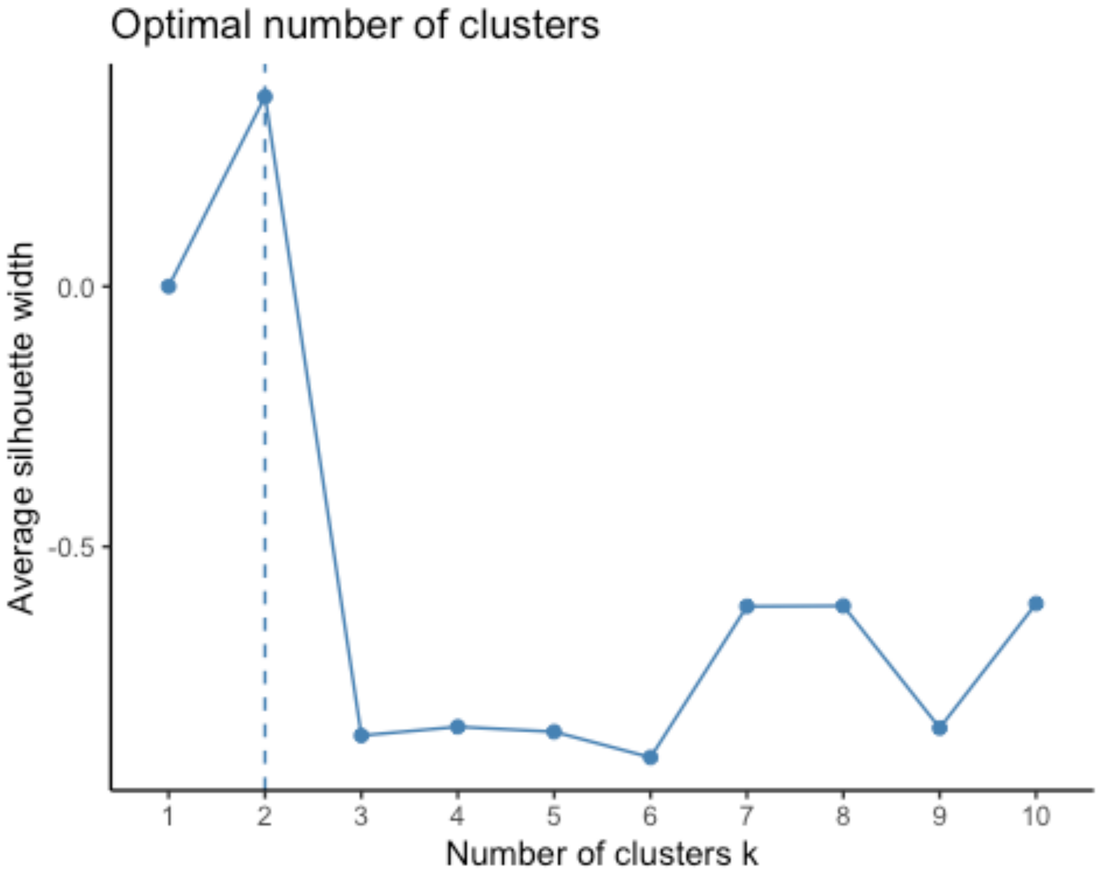






Filling NA Values with PreProcessing

I filled in the NA values and use preprocess() from caret package to impute the missing data. I can use three methods for imputation: knn, bag, median. I won't generate dummy variables for scale data because even though the dataset description has noted them as scaled some data points are float variables. According to clara(like kmeans) clustering 2 per group is an optimal number for clusters, I used clara instead of kmeans because there are many NA values in the dataset.



Although k=2 clustering for knn or bag(like random forest but with smaller numbers of trees) imputation are optimal, the data set has a large number of NA values so for pre-processing, I am going to conduct a two sample t-test on the means of total NA values in each developed and developing country entry to find out if the NA values are randomly distributed amongst the two different groups of countries. If randomly distributed than I would be able to use a median impute method to impute the missing values with the sample medians since the assumption of median impute method relies on the NA values being distributed randomly in the data set in order to avoid biased sampling during my analysis.

Two Sample T-Test for Mean Analysis of Random NA distribution

```
#H0=na values are randomly distributed across developing and developed countr
ies, x_bar1=x_bar2
#HA=na values are not randomly distributed across developing and developed co
untries, x_bar1!=x_bar2


##
##  Welch Two Sample t-test
##
## data:  na_sums by y
## t = -0.28272, df = 283.59, p-value = 0.7776
## alternative hypothesis: true difference in means between group 0 and group
```

```
1 is not equal to 0
## 95 percent confidence interval:
##  -33.27120  24.91405
## sample estimates:
## mean in group 0 mean in group 1
##        125.2500        129.4286
```

According to the hypothesis test NA values are randomly distributed as we do not reject the null hypothesis since the p-value is much larger than alpha=0.05. Although the median impute method is not the ideal pre processing strategy, for example knn or bag imputations would be much preferable for reasonable predictions, with my large amount of missing data I have to utilize this convenient and optimal method which I can satisfy its assumptions within my data. In the analysis of this data mining project, I will account for the statistical imbalances or misinterpretations this median imputing method could cause. According to the steps of the process of median imputation above; it ignored 1 variables and imputed data for all variables. I will now use this model to predict the missing values in OECD_21c.

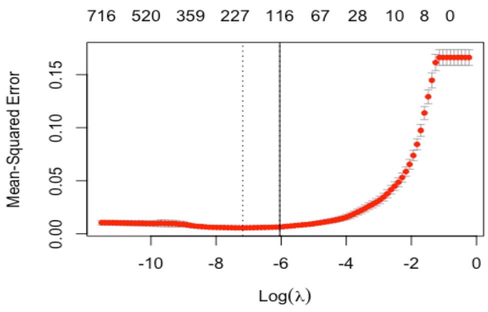
Data Mining

In this part I will run models and look into veiled patterns within the data. My goal in this data mining project is to determine the most significant indicators that indicate the development level of OECD countries. I will be able to determine the most significant features or indicators by collecting the most frequently identified indicators by the models I train my data on. I will first start with Lasso because it will assign a coefficient value of zero to insignificant indicators and I will be able to identify the number of explanatory/predictor variables necessary for my other models. After receiving the list of significant indicators and the amount of them, I will select the same number of indicators that are most significant in my other models which are logistics regression, pca+logistic regression, NaiveBayes, and RandomForest.

Lasso(standardized)

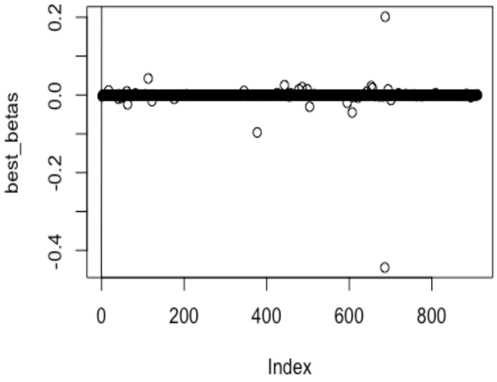
Sets coeff to absolute zero if not significant so a Lasso model is a good method for feature selection within the model.

```
plot(lasso.cv3)
abline(v=log(lasso.cv3$lambda.1se))
```



```
#best lambdas are in cv3 so we move forward with that model:
best_lambda <- which(lasso.cv3$lambda == lasso.cv3$lambda.1se)
```

```
best_betas <- lasso.cv3$glmnet.fit$beta[, best_lambda]
plot(best_betas)
abline(a=0,b=1)
```



```
mean(best_betas==0)

## [1] 0.8726674

## Confusion Matrix and Statistics
##
##          actual
## predicted   0   1
##         0  40   0
##         1   3 115
##
##                Accuracy : 0.981
##                  95% CI : (0.9455, 0.9961)
##     No Information Rate : 0.7278
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.951
##
##  Mcnemar's Test P-Value : 0.2482
##
##             Sensitivity : 1.0000
##             Specificity : 0.9302
##          Pos Pred Value : 0.9746
##          Neg Pred Value : 1.0000
##              Prevalence : 0.7278
##          Detection Rate : 0.7278
##    Detection Prevalence : 0.7468
##       Balanced Accuracy : 0.9651
##
##        'Positive' Class : 1
##
```

```
##   model classification_error precision recall sensitivity
## 1 lasso           0.01898734 0.9745763      1           1
```

Looks like a good fit model! But when we look at the confusion matrix data detection rate is low which means that measure ability to actually detect the diverse groups is low. Although the four metrics we use to evaluate the efficiency of the model are perfect, low detection rate does indicate that this model might not be the best fitting model. Yet, we can conclude that the top features selected by this model can be the most significant features or indicators on the development level of a country and the number of significant predictor variables selected is the ideal.

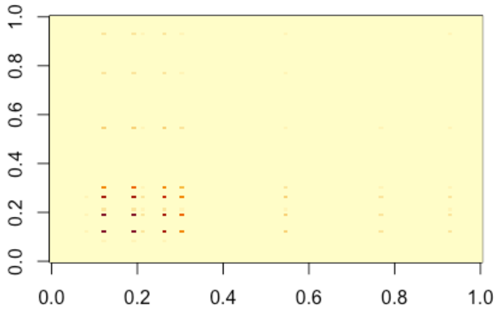
Since the Lasso model is very good at selecting the significant coefficients, I will look into all of them and utilize them to compare with the other models' selected indicators/features.

Logistic Regression with All Features (no scaling nor tuning-raw model)

I will select top 116 indicators to be able to compare the most significant 116 indicators with the other models and try to come up with a conclusion on what indicators contribute mostly to the development of an economy, I picked 116 because Lasso determined 116 indicators to be significant

```
## Confusion Matrix and Statistics
## 
##          actual
## predicted  0  1
##         0 20 63
##         1 23 52
## 
##                Accuracy : 0.4557
##                  95% CI : (0.3764, 0.5367)
##     No Information Rate : 0.7278
##     P-Value [Acc > NIR] : 1
## 
##                   Kappa : -0.0641
## 
##  Mcnemar's Test P-Value : 2.605e-05
## 
##             Sensitivity : 0.4522
##             Specificity : 0.4651
##          Pos Pred Value : 0.6933
##          Neg Pred Value : 0.2410
##              Prevalence : 0.7278
##          Detection Rate : 0.3291
##    Detection Prevalence : 0.4747
##       Balanced Accuracy : 0.4586
## 
##        'Positive' Class : 1
## 
```

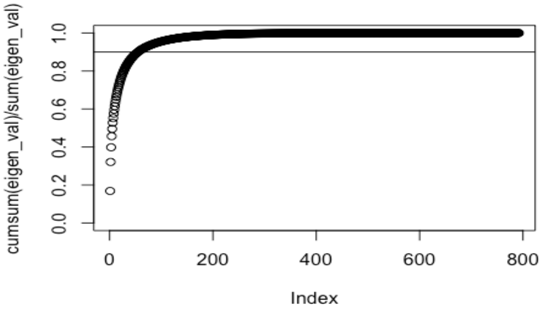
```
image(cov(sample(trainData, 100)))#it is hard to visualize collinearity among all data so when we sample 100 we can see that there is collinearity in the d
ata and thus we should utilize PCA to overcome this handicap
```



In this model of logistic regression where all features are included the model summary shows us that none of the beta values are significant. This might be because there are too many predictor variables that are collinear or interacting intrinsically with each other. The metrics suggest that the model fits poorly. The model is predicting less accurately and the predictors have less significance compared to previous model which can also be a repercussion of the median imputation method I used. For a better model, we need feature selection so I will use PCA and logistic regression together next to determine the significant features.

PCA + Logistic Regression for Feature Selection and Model Fitting(scaled)

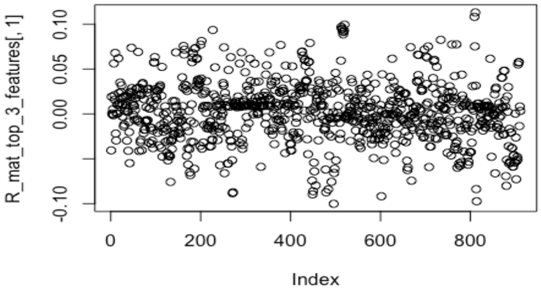
I will use PCA to move forward in selecting the most significant features.



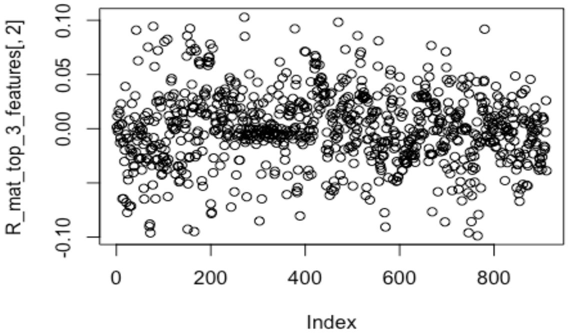
*I pick the k to be 50 because the kink happens at that point and the variability is lower after 50

PCA+Logistic Regression

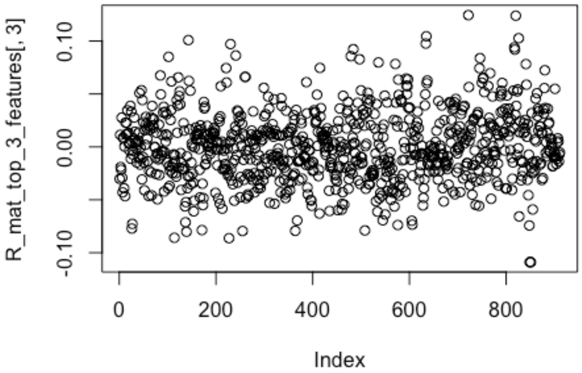
```
plot(R_mat_top_3_features[,1])
```



```
plot(R_mat_top_3_features[,2])
```



```
plot(R_mat_top_3_features[,3])
```



I will take loadings values that are above 0.02 based on the plots I visualized and only get top 2 because the third plot loadings don't provide significant information

```
## Confusion Matrix and Statistics
##
##           actual
## predicted   0   1
##         0  41   1
##         1   2 114
##                                           
##                Accuracy : 0.981           
##                  95% CI : (0.9455, 0.9961)
##     No Information Rate : 0.7278          
##     P-Value [Acc > NIR] : <2e-16          
##                                           
##                   Kappa : 0.9517          
##                                           
##  Mcnemar's Test P-Value : 1               
##                                           
##             Sensitivity : 0.9913          
##             Specificity : 0.9535          
##          Pos Pred Value : 0.9828          
##          Neg Pred Value : 0.9762          
##              Prevalence : 0.7278          
##          Detection Rate : 0.7215          
##    Detection Prevalence : 0.7342          
##       Balanced Accuracy : 0.9724          
##                                           
##        'Positive' Class : 1               
##                                           

##                        model classification_error precision    recall
## 1 pca + logistic regression           0.01898734 0.9827586 0.9913043
##   sensitivity
## 1   0.9913043
```

The PCA + logistic regression can detect a signal between our “W” and Y and it has a lower classification error compared to logistic regression model. The model has selected 232 features to be indicative of development of a country which is more than the number of indicators lasso had selected. The model can’t avoid random noise as we get the necessary features list to be very long. I will try out other algorithms to compare which is the best fitting.

NaiveBayes(tuned)

```
## Confusion Matrix and Statistics
## 
##       prediction
## actual   0   1
##      0  34   9
##      1   0 115
##                                           
##                Accuracy : 0.943           
##                  95% CI : (0.8946, 0.9736)
##     No Information Rate : 0.7848          
##     P-Value [Acc > NIR] : 3.518e-08       
##                                           
```

```
##                    Kappa : 0.8461
##                                   
##  Mcnemar's Test P-Value : 0.007661
##                                   
##             Sensitivity : 1.0000
##             Specificity : 0.9274
##          Pos Pred Value : 0.7907
##          Neg Pred Value : 1.0000
##              Prevalence : 0.2152
##          Detection Rate : 0.2152
##    Detection Prevalence : 0.2722
##       Balanced Accuracy : 0.9637
##                                   
##        'Positive' Class : 0
##
```

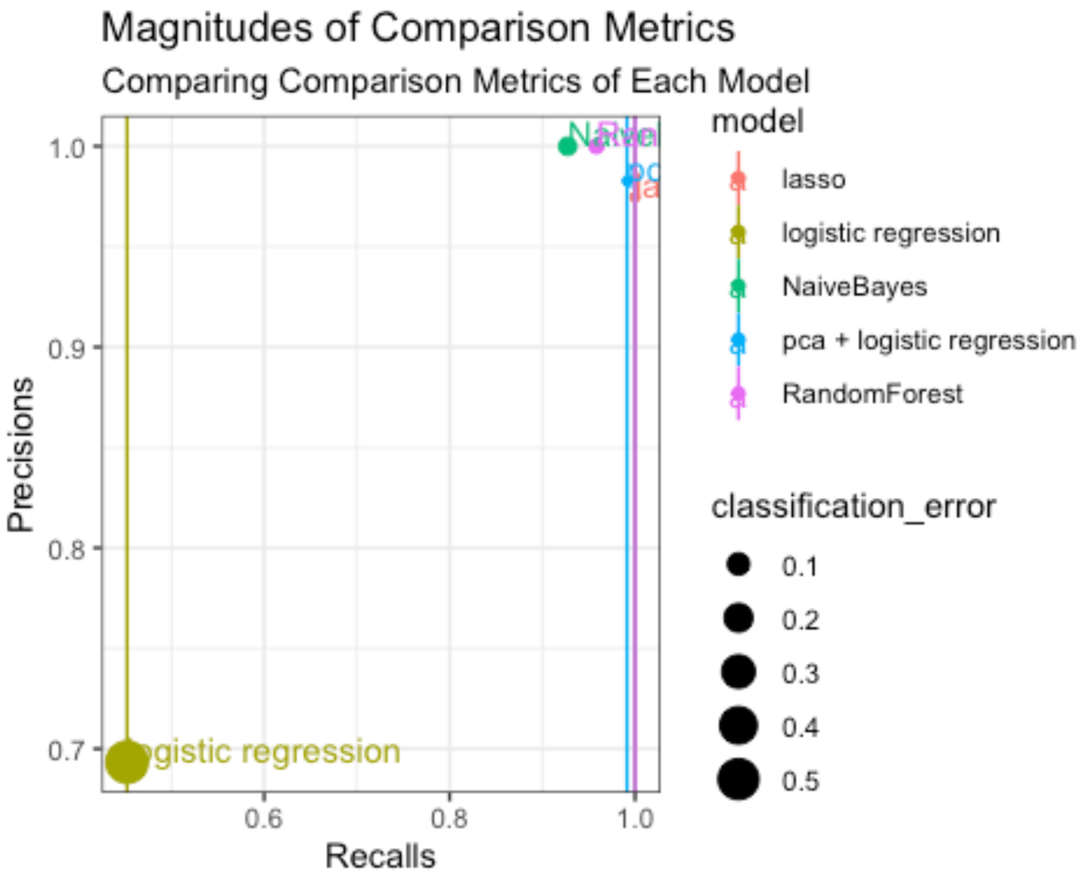
```
##        model classification_error precision    recall sensitivity
## 1 NaiveBayes           0.05696203         1 0.9274194           1
```

RandomForest(tuned)

```
## Confusion Matrix and Statistics
## 
##       prediction
## actual   0   1
##      0  38   5
##      1   0 115
##                                           
##                Accuracy : 0.9684
##                  95% CI : (0.9277, 0.9896)
##     No Information Rate : 0.7595
##     P-Value [Acc > NIR] : 3.615e-13
##                                           
##                   Kappa : 0.9171
##                                           
##  Mcnemar's Test P-Value : 0.07364
##                                           
##             Sensitivity : 1.0000
##             Specificity : 0.9583
##          Pos Pred Value : 0.8837
##          Neg Pred Value : 1.0000
##              Prevalence : 0.2405
##          Detection Rate : 0.2405
##    Detection Prevalence : 0.2722
##       Balanced Accuracy : 0.9792
##                                           
##        'Positive' Class : 0
##
```

```
##          model classification_error precision    recall sensitivity
## 1 RandomForest           0.03164557         1 0.9583333           1
```

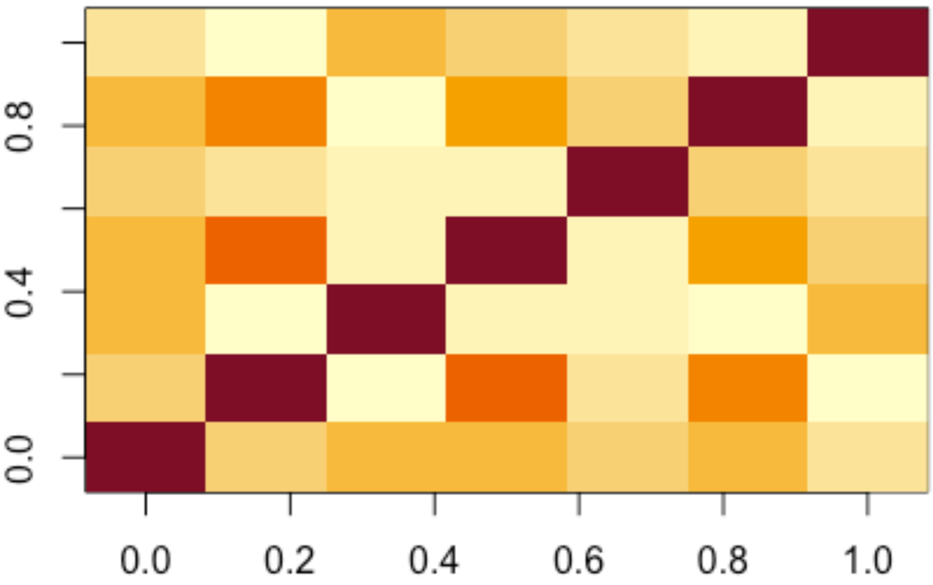
Comparing the Metrics

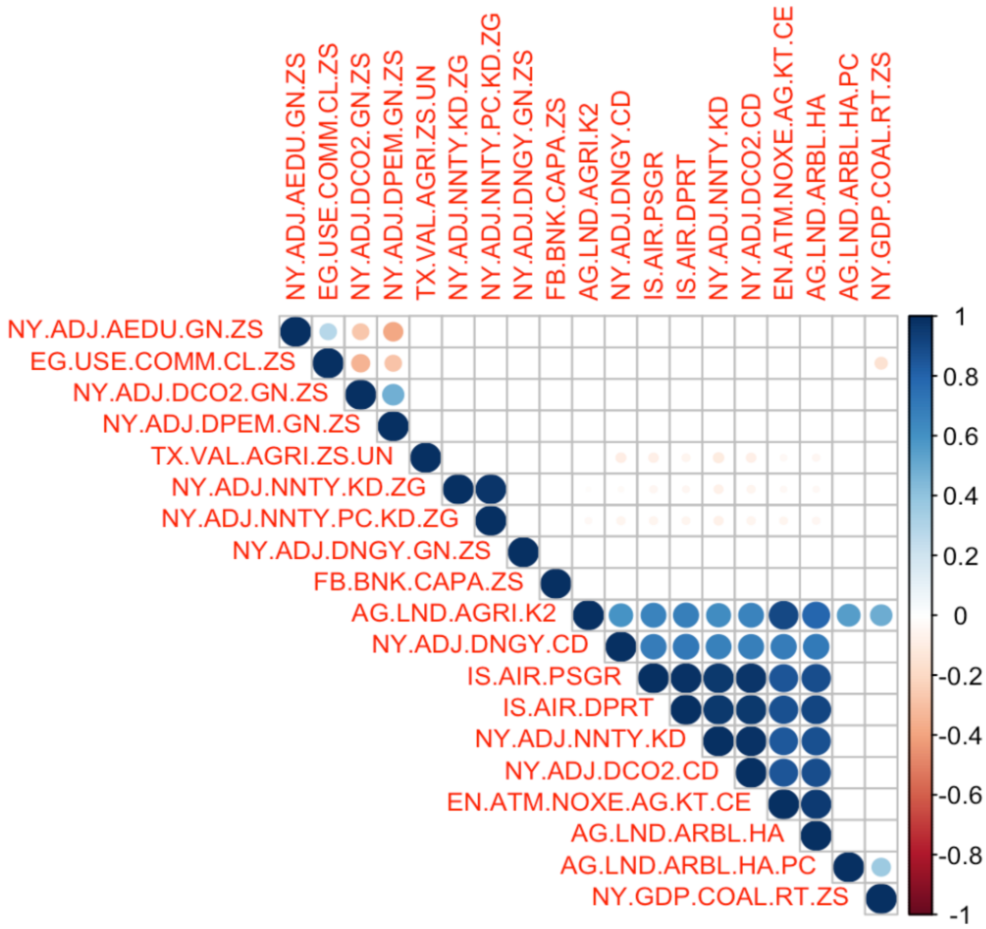


In this metrics comparison graph we see that scaled lasso and tuned random forest and naive bayes have the same sensitivity which is 1. So for creating the best fit model with significant indicators from the financial crisis year 2008 I will get the significant indicators determined by the lasso model and fit my model with random forest because lasso has highest recall and random forest has highest precision with less classification error. I will also engineer a new feature utilizing the significant indicators that were selected by all the models and include it in this final model. To asses the significance of my engineered feature I will conduct a chi-square test using anova() for it by fitting a logistic regression with the most significant indicators that were in all models.

Feature Engineering

```
image(joint_feature_corr_matrix)#there is correlation among joint features
```


0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0



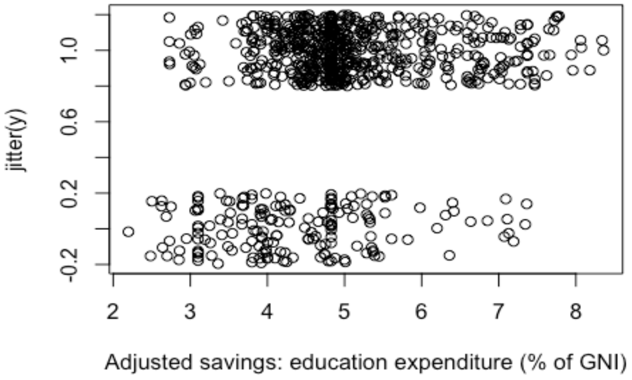
Based on the most significant correlations plot above, “IS.AIR.PSGR”(Air transport, passengers carried), “IS.AIR.DPRT”(Air transport, registered carrier departures worldwide), “NY.ADJ.NNTY.KD”(Adjusted net national income (constant 2015 US)), and “NY.ADJ.DCO2.CD”(Adjusted savings: carbon dioxide damage (current US)) are all highly positively correlated.

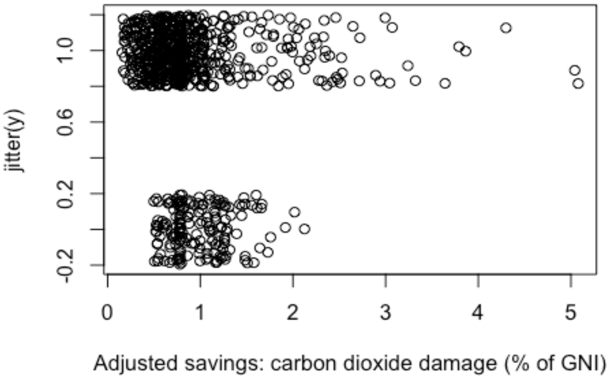
“NY.ADJ.NNTY.KD.ZG”(Adjusted net national income (annual % growth)) is highly positively correlated with “NY.ADJ.NNTY.PC.KD.ZG”. (Adjusted net national income per capita (annual % growth))

“AG.LND.AGRI.K2”(Agricultural land (sq. km)) is highly positively correlated with “EN.ATM.NOXE.AG.KT.CE”(Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)).

“NY.ADJ.AEDU.GN.ZS”(Adjusted savings: education expenditure (% of GNI)) is negatively significantly correlated with “NY.ADJ.DPEM.GN.ZS”(Adjusted savings: particulate emission damage (% of GNI)) and “NY.ADJ.DCO2.GN.ZS”(Adjusted savings: carbon dioxide damage (% of GNI))

One of the most interesting correlation among the most significant indicators, in my opinion, is the negative correlation betwee education expenditure and cost of carbon dioxide damage. Below are the visualizations:





The graphs indicate that for developed countries the rate of education expenditure is high and cost of carbon dioxide damage is low while it is the opposite in developing countries. The visualizations don't provide explicit data on effect on education and especially female education so I am going to engineer a feature based on female education indicators to account for the effect of education among the most significant indicators. The importance of education is obvious from this analysis and could be represented as another variable.

```
#from the indicators that specify data about female education and schooling I am selecting the ones below:
education_ind_names <- c("Share of youth not in education, employment or training, female (% of female youth population)", "Primary education, pupils (% female)", "Labor force with advanced education, female (% of female working-age population with advanced education)", "Compulsory education, duration (years)")
#I will normalize between 0 and 1 each indicator(except compulsory education years) then I will multiply all indicator data to generate a females education score that will be added to the best fit model I will generate

#Now lets test the significance of this index I engineered with anova():
#H0=the female education index I engineered is not significant--coefficient=0
#HA=the female education index I engineered is significant--coefficient!=0
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ (NY.ADJ.DCO2.GN.ZS + AG.LND.AGRI.K2 + TM.VAL.AGRI.ZS.UN +
##     AG.LND.ARBL.HA.PC + FB.BNK.CAPA.ZS + NY.GDP.COAL.RT.ZS +
##     EG.USE.CRNW.ZS + female_education_index) - female_education_index
## Model 2: y ~ NY.ADJ.DCO2.GN.ZS + AG.LND.AGRI.K2 + TM.VAL.AGRI.ZS.UN +
##     AG.LND.ARBL.HA.PC + FB.BNK.CAPA.ZS + NY.GDP.COAL.RT.ZS +
##     EG.USE.CRNW.ZS + female_education_index
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       784     625.49
## 2       783     491.60  1    133.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the hypothesis test the female education index is significant for the model that was created with the indicators that were selected to be significant by all models in this data mining project. The p-value of the chi-square test of significance is less than alpha=0.05 so we reject the null hypothesis. The female education index captures the accessibility and quality of female education as it entails data on females that are not taking part in any daily activity (an indicator of unfairness towards females), females access to primary education, educated females proportion in the labor force, and compulsory education duration which enables female students to attend school by law.

Final Best Fit Models Comparison

As I have also stated before, I am going to run the two final RandomForest models to assess the significance of 2008.

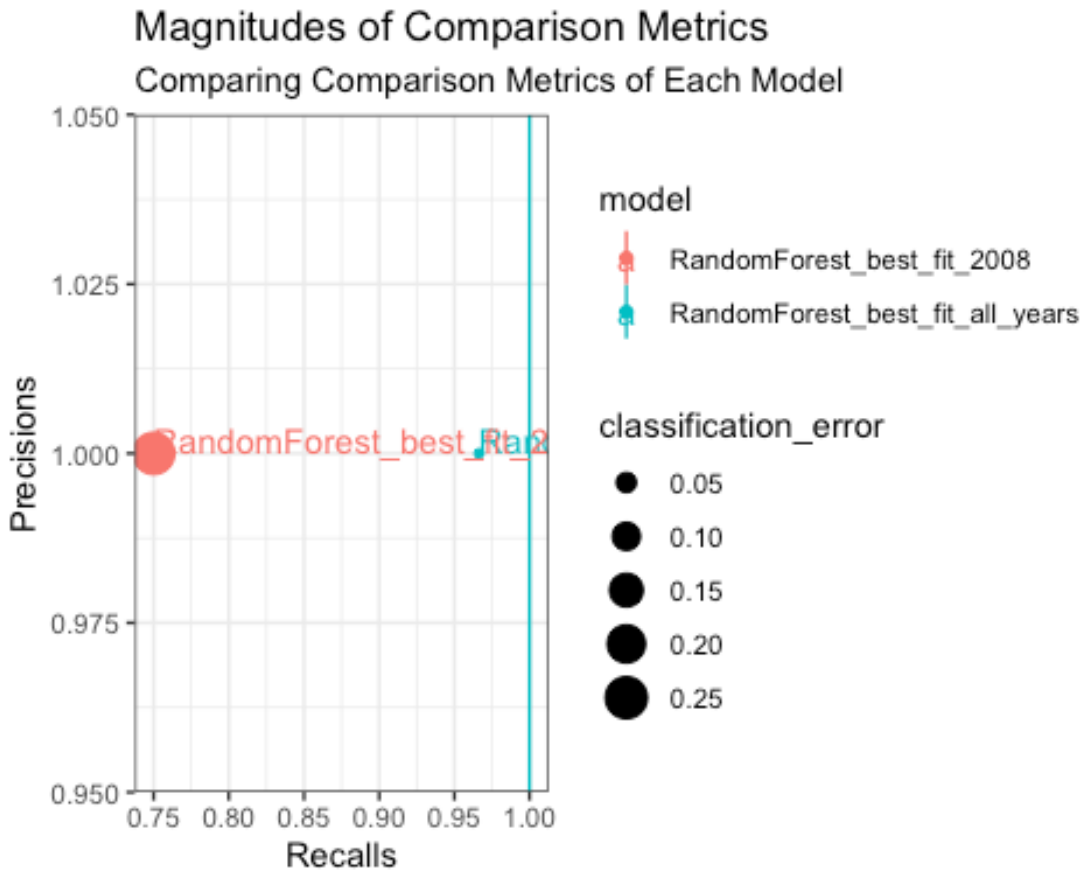
predictor variables = female education index and 116 significant features selected by lasso model only for year 2008

```
## Confusion Matrix and Statistics
##
##       prediction
## actual 0 1
##      0 0 2
##      1 0 6
##
##                Accuracy : 0.75
##                  95% CI : (0.3491, 0.9681)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1.0000
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.4795
##
##             Sensitivity :   NA
##             Specificity : 0.75
##          Pos Pred Value :   NA
##          Neg Pred Value :   NA
##              Prevalence : 0.00
##          Detection Rate : 0.00
##    Detection Prevalence : 0.25
##       Balanced Accuracy :   NA
##
##        'Positive' Class : 0
##

##                         model classification_error precision recall sensiti
vity
## 1 RandomForest_best_fit_2008                 0.25         1   0.75
NA
```

Let’s also try RandomForest without year specification since the classification error is high for only year 2008:

```
## Confusion Matrix and Statistics
##
##       prediction
## actual   0   1
##      0  39   4
##      1   0 115
##
##                Accuracy : 0.9747
##                  95% CI : (0.9365, 0.9931)
##     No Information Rate : 0.7532
##     P-Value [Acc > NIR] : 1.105e-14
##
##                   Kappa : 0.9342
##
##  Mcnemar's Test P-Value : 0.1336
##
##             Sensitivity : 1.0000
##             Specificity : 0.9664
##          Pos Pred Value : 0.9070
##          Neg Pred Value : 1.0000
##              Prevalence : 0.2468
##          Detection Rate : 0.2468
##    Detection Prevalence : 0.2722
##       Balanced Accuracy : 0.9832
##
##        'Positive' Class : 0
##

##                              model classification_error precision    recall
## 1 RandomForest_best_fit_all_years           0.02531646         1 0.9663866
##   sensitivity
## 1           1
```

Conclusion



I was expecting to get a better fit with specifying a crisis year because I thought the model would be a better fit if it is setup according to a specific year with specific attributes that minimize the unaccounted exogenous variables. In the two RandomForest models I fitted, I realized that the one that has data accounting for all years without a year explanatory variable is a better fitted model. I had initially thought that year had to be a significant variable that accounts for major historical, technological and political events yet after comparing the final models and the previous other models, I realized that year is not significant. Year wasn't selected by all models either since it doesn't come up when I filter all significant indicators from all models to find the overlapping indicators.

In conclusion, organizations that are interested in calculating indicators that significantly imply or address development levels of countries and the effect of each indicator on future development should focus primarily on the joint_indicators I got from all models which are:

[1] "Adjusted net national income (annual % growth)"
[2] "Adjusted net national income (constant 2015 US$)"

[3] "Adjusted net national income per capita (annual % growth)"

[4] "Adjusted savings: carbon dioxide damage (% of GNI)"

[5] "Adjusted savings: carbon dioxide damage (current US$)”
[6] “Adjusted savings: education expenditure (% of GNI)”
[7] “Adjusted savings: energy depletion (% of GNI)”
[8] “Adjusted savings: energy depletion (current US$)”
[9] “Adjusted savings: particulate emission damage (% of GNI)”
[10] “Agricultural land (sq. km)”
[11] “Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)”

[12] “Agricultural raw materials exports (% of merchandise exports)”
[13] “Air transport, passengers carried”
[14] “Air transport, registered carrier departures worldwide”
[15] “Alternative and nuclear energy (% of total energy use)”
[16] “Arable land (hectares per person)”
[17] “Arable land (hectares)”
[18] “Bank capital to assets ratio (%)”
[19] “Coal rents (% of GDP)”

This list can be expanded to a total of 117 variables including the feature I generated (female education index) and the significant variables lasso model has detected. For countries that weren’t included in this dataset, year or other politic indicators could be significant as these other countries won’t be part of a group like OECD. The ideal dataset would be the data set where no missing values exist because in this data mining project, I had to impute data which might have affected the overall performance of all models. We should also consider the positive effect the median imputation method could have caused because the metrics of the models I used were close to each other but with real data the metrics data could have been more diverse.

All the findings of this project provide further information for countries as it specifies the most important indicators that help determine development level. Also, the model can be used to fit specific country or groups of country data to measure development levels by year or by country. Organizations like the IMF, World Bank, or the UN would utilize these findings in their aid programs as they would be able to determine what type of funding countries need to develop themselves more. Further research on this topic should be conducted if datasets without missing data can be acquired since this project worked on data that was imputed. The imputation might have caused some biases so the effect should be kept in mind during regeneration of these models. For robust results having large datasets is important. The positive effect of having a large dataset was obvious during the last part where I fitted RandomForest on all years and only 2008. The model with all years was a better fit compared to the small dataset with only 2008 data so large dataset is an important key to success. Uncertainty is an issue that arises with large datasets, in such cases median imputation is a method that can be a convenient remedy which I have also utilized. If median imputation is implemented then its effects should also be acknowledged during the analysis as I have done.

Feedback on Ohad and Joseph’s Project

IRS Data – NonProfit Organization Project

I loved your idea and the topic you are looking into. I think the dataset you selected is also very comprehensive to understand the factors that support successful non-profit organizations. I find it difficult to work with categorical explanatory variables but the skewness analysis and the data wrangling you have done all foreseeable obstacles you could come across seem to be tackled. In the data exploration and cleaning part I realized that you assigned zero to NA values. I also had. A lot of NA values and. Used median imputation to resolve the issue, why did you prefer this method? Did it affect your data analysis in any way? My median imputation method has overfitted my data for example and I tried to account for it in my analysis. In your data mining part, I see that you have done feature selection with Lasso. I also preferred the same method for feature selection I think it is very effective and convenient. Overall, I think the findings of this project would benefit many organizations and consequently help a lot of people.