Group 3 – Final Report

Ivan Escalona-Faria
Computer Science Student
Georgia State University
Atlanta, United States
iescalonafaria1@student.gsu.edu

***Abstract*— The following work contains the report of the final project of Big Data Programming (CSC-4760) corresponding to the fall 2019 semester. Random Forest Classifier and Regressor were used with economic indicators data from the World Bank in order to predict the Human Development Index (HDI) for a country, as well as the GINI Coefficient. The accuracy of the models was were above 96% in all cases. This information can be used to help in reduce the dimensionality of the almost 1600 indicators, as well as discover factors that affect HDI and GINI Coefficient in order to focus efforts on these to improve these scores for underperforming countries.**

Keywords—component, formatting, style, styling, insert (*key words****)***

I. Introduction

The economy is one of those aspects of life in which everyone has a vested interest. Having finite resources and typically more needs than resources, the economy plays a major role in both explaining as well as improving the wellbeing of individuals all the way to nations and the whole world.

It is a very complex field, with multiple interdependent variables – and even that's when the data is available – that are constantly changing, very often in unpredictable patterns. Out of the many economic indicators that relate to the overall health of the economy of a nation, there are two such indicators that are often good in pointing whether a nation is on the right track or not, these are: Human Development Index (HDI), and the GINI Coefficient.

The first one relates to how developed a nation is, this is measured from 0.0 to 1.0 with 1 being completely developed[1]. Whereas the GINI Coefficient is a measure of the income inequality in a nation, where 0 represents complete equality in the distribution of income and 1 represents complete inequality[2].

The World Bank hosts a database of about 1600 different economic indicators that range from 1960 to 2017 for every country, and even aggregate regions in the world. This rich data set can be mined to extract interesting insights regarding these two values. The idea in principle is to identify the features that have a greater impact (from the model's perspective) in determining the HDI for a given country on a given year, and once this is determined, if a similar model is able to predict the GINI Coefficient given a set of values for the rest of the indicators. These two can prove very valuable to help improve the situation of countries in which there is great income inequality and are having a hard time improve their development state.

Two questions were asked with the hope of answering them by using Data Mining/Machine Learning techniques:

Question 1: Which factors are the most important in explaining the development of countries (As per the Human Development Index)?

Question 2: Can it be accurately predicted which countries saw an increase in income inequality for 2018? (using GINI Coefficient)?

II. Pre-Processing

The analysis was performed by using Python 3, and the some of the typical Python modules such as Numpy, Pandas, Matplotlib, Seaborn, and Scikit Learn.

The first problem that was faced while exploring the data was its structure. The columns are reversed with the features and there are a great number of rows (422,136), but these don't correspond to a given country-year pair for all features, but for each feature. The first thing was to transpose this and develop a structure in which the economic indicators are columns and that the rows (instances) correspond to a given country for a corresponding year.

Even before this, given the great amount of data that is present, particularly with regard to the number of features, by performing a groupby and counting for every country how many rows are there (before transposing the data frame) and then determining the number of unique features that are present in this, it was revealed that the only value is 1599. This is an assurance that there are 1599 features for all countries, even if they are null values. This is important to

ensure the consistency of the Data Frame in order to perform operations on it. It is also seen that the values of all the features are numerical, this is important as the only thing to perform on them should be a normalization right before train-test splitting.

The next major challenge, as per usual, is the existence of null values in the Data Frame. The first thing was to remove all columns that had only null values. After this there was an attempt to remove all rows that had at least one null value, but this gives an empty data frame. Since there are so many features, the probabilities of having at least one null value in one of the fields is very high. One way to overcome this was to try and set a threshold to remove only columns with a high number of null values (the threshold was more than 80% null values), and then rows with more than 35% null values. This reduces the data from 1599 to 1044 columns and from 15392 to 4687 rows. This could be one approach to be used, the other one could be to leave them as is, and imputate the values by filling the null values with a given value – for instance the mean of the column. This however is not typically recommended as this might skew the data toward these values, further more there might be features in which the mean won't make sense and a domain expert will have to make the assessment on a case by case basis.

For example, if the GDP was to be averaged, there might be a situation in which there is a missing value for a country in which by its size, the GDP will not be nearly the world's average GDP. However to treat this will be cumbersome, time consuming, and will require a domain expert. So the risk is acknowledged, and the analysis will proceed accordingly.

The next step is to incorporate the Human Development Index by country and year into the DataFrame. The operation should be actually performed on the full Data Frame (data) since the HDI data only encompasses years 1990 onward, so 30 years after worth of information will be lost from the Data Frame. It is a safer approach to drop the null values after this merger occurs. After this, the HDI data set was merged and encoded with values 0,1,2 which according to the website tutor2u.net[4], a low (<0.5), medium (between 0.5 and less than 0.8), high (>0.8) range can be established for the HDI, this approach was preferred since it reduces the number of labels from a continuous to only 3. The less labels the better the behavior of the model.

A Max-Min normalization (Range 0 to 1) was performed on all of the numeric features. This was preferred over a Z-Score standardization since it is hard to assess whether the data has a normal distribution or not.

III. Machine Learning/Data Mining Analysis

The data shows a significant imbalance for each of the three targets (0,1,2), 61%, 20%, and 19% respectively. This suggests the need for stratified splitting. The size of the split was 90% training, 10% test.

A random forest classifier was used to model the data with the ultimate goal of extracting the quantified order of importance of each of the 1044 features entered as input in the training. The information gain criteria used were both Gini and Entropy respectively, the number of estimators were 100, and the Max. depth allowed was also changed from 1 to 15. The rest of the parameters were left as per the default ones in the module. The results are summarized on Table 1 as well as figure 1 below:

Table 1. Random Forest Classifier results

Criterion	*Max. Depth*	*Train Score*	*Test Score*
Gini Index	1	0.759	0.757
	2	0.870	0.860
	3	0.903	0.895
	4	0.926	0.919
	5	0.951	0.941
	6	0.966	0.950
	7	0.980	0.967
	8	0.988	0.972
	9	0.991	0.967
	10	0.994	0.978
	11	0.996	0.976
	12	0.998	0.985
	13	0.999	0.985
	14	0.999	0.982
	15	0.999	0.985
Entropy	1	0.748	0.751
	2	0.884	0.880
	3	0.905	0.897
	4	0.930	0.928
	5	0.956	0.945
	6	0.976	0.956
	7	0.987	0.969
	8	0.994	0.974
	9	0.998	0.978
	10	1.000	0.980
	11	1.000	0.982
	12	1.000	0.980
	13	1.000	0.982
	14	1.000	0.982

Table 1. Random Forest Classifier results			
Criterion	***Max. Depth***	***Train Score***	***Test Score***
	15	1.000	0.980

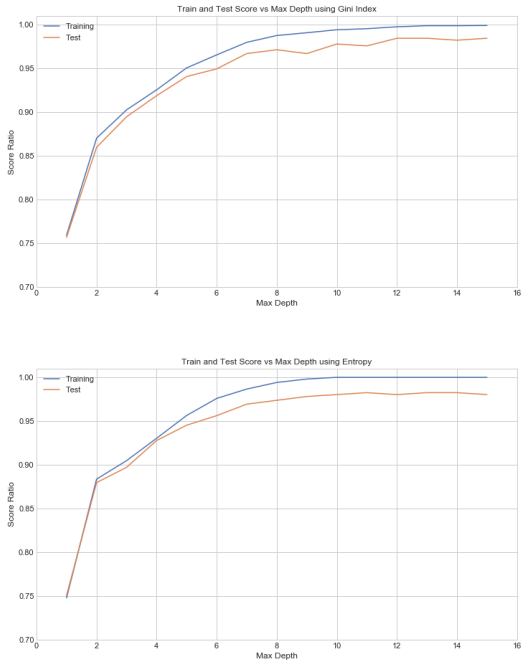

Figure 1. Train and test score vs Max. Depth. (a) Gini index (b) Entropy

The values obtained were good, certain considerations are to be made. In the case of the Gini-Random Forests there is some erratic behavior as the Max. depth increases, the best tree would be at depth 15, however this erratic behavior increases the risk of memorization of the model as opposed to generalization. It's better to rely on the Entropy model, but at max. depth 10 the model reaches 1.00 on the training accuracy. This is a dangerous point, as the perfect match increases the risks of memorization, so the best option is to take the previous one, which still provides a test score of 0.978.

Other metrics such as the classification report and the confusion matrix are shown below in table 2 and figure 2 respectively.

Table 2. Classification report				
	f1-score	**precision**	**recall**	**support**
0	0.977	0.977	0.977	87
1	0.982	0.975	0.989	278
2	0.967	0.989	0.946	92
accuracy	0.978	0.978	0.978	0.978
macro avg	0.975	0.980	0.971	457
weighted avg	0.978	0.978	0.978	457

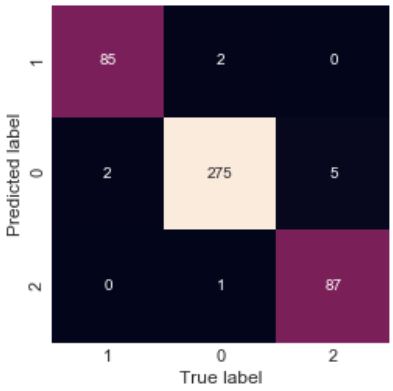

Figure 2. Confusion Matrix.

These continue to confirm good results for the model. The features were ranked in order of importance. They are too many to be shown here, but an aggregate of these is displayed on table 3.

Table 3. Feature Importance Aggregate	
Number of Features	***Accumulated Feature Importance***
100	0.692
200	0.799
300	0.855
400	0.894
500	0.923
600	0.947
700	0.965
800	0.979
900	0.988
1000	0.995
1100	0.999
1200	1.000

A good insight comes from this table. With only 300 features, more than 85% of the data can be classified. This

would allow for a dimensionality reduction of more than 5 times what the original data had. The top 10 features are shown on Table 4.

GINI Coefficient prediction

On the GINI Coefficient prediction, the first 300 features in importance from the previous model, this allows the computation to be significantly faster and focus on the more relevant features as per the model. The same split size and model parameters as with the best forest for the previous model were used.

The results are equally encouraging as the model is still showing an accuracy of more than 98%, and for specific data on 2017 it showed more than 96% accuracy.

IV. Analysis

The top 10 important features were extracted. They only add up to 20% of the data but can provide some idea as to what further relevance or usage they can have at the time of extracting insights about the domain of nations development and income inequality. These are summarized on table 4:

Table 3. Feature Importance Aggregate

Rank	*Feature name*	*Ratio of importance*
1	*Mortality rate, neonatal (per 1,000 live births)*	*0.027874266*
2	*Mortality rate, adult, male (per 1,000 male adults)*	*0.024642132*
3	*Adjusted net national income per capita (constant 2010 US$)*	*0.022163601*
4	*Households and NPISHs Final consumption expenditure (current US$)*	*0.020255716*
5	*Lending interest rate (%)*	*0.020120948*
6	*Mortality rate, infant (per 1,000 live births)*	*0.019636265*
7	*GDP growth (annual %)*	*0.018612188*
8	*Access to clean fuels and technologies for cooking (% of population)*	*0.018195125*
9	*GNI per capita, PPP (constant 2011 international $)*	*0.018028401*
10	*GDP per capita (current LCU)*	*0.016980809*

From this it can be said that mortality plays an important role in determining the HDI, as well as some GDP, and PPP values. It can't be said whether the relationship is direct or inverse, however this conveys more specific details as to which mortality rate is the most important, and the same can be said about the GDP and PPP indicators. The point can be made then, that in order to develop a country, special attention must be paid to the mortality rate. These are of course somewhat trivial and intuitive, as in the very same definition of HDI this is reflected.

However, an interesting feature comes ranked in 5th place, that is the "Lending interest rate (%)", this is rarely thought of as an indication of development of a nation, yet it is in the top 10. An even more palpable one is ranked 8th "Access to clean fuels and technologies for cooking (% of population)". It is fair to say that this would not be the first thing that comes to mind as a determining factor when classifying a nation as developed or underdeveloped.

These can become important insights if taken into consideration at the time of invest in efforts to enhance the development of nations. It could be the case that a country with low lending interest rates will have better chances to secure investments and stimulate the economy altogether rather than blindly targeting other factors.

Regarding the GINI Index, it is ranked 976 out of all the factors taken into the Random Forest Classifier. With 0.0059% this may indicate that, it is a factor in deciding whether a nation is developed or not, but there are at least 976 other factors with more weight than the GINI Index. So as it turns out the income inequality might not play as big of a role as was believed.

The prediction of the GINI Index was good even for the single year (2017) features. This will help direct the efforts toward improving the income inequality, because this means that, at least for the moment, it accurately predicts future GINI Coefficients, so provided the distributions continue to be stable, this can be used to try and model future GINI Coefficient values based on changes or predictions on other indicators and therefore make necessary corrections on time, that is, before it has already happened.

V. Conclusions

A Random Forest Classifier and Regressor were used to fit the economic indicators from the World Bank data set. The most prominent features were extracted in order to reduce the dimensionality of the data, as well as provide insights on the more important factors that determine the HDI for nations. This provided interesting and novel insights on important features that contribute to determining the HDI and therefore the development of a country, such as mortality rates, GDP and PPP indicators, but also lending interest rates and access to clean fuels and technologies for cooking.

This was also done for the prediction of the GINI Coefficient. All models provided an accuracy of more than 96% and it is presumed that no memorization occurred for these models.