Data Exploration and Analysis of the CIA World Fact book

Noel Carrascal
noelcarrascal@gmail.com

December 15, 2015

Introduction

The CIA World Fact book provides information on every country and territory in the world. The information is gathered and published yearly by the CIA. Statistics on each country are collected for 30 fields that are grouped in five different categories: People and Society, Geography, Military, Communications and transportation, and from 2004 to 2014.

The data are processed to eliminate null values. For some fields that are null, the data is imputed using a linear regression, and in a few important cases manually. After cleaning the data of null values, only seventy seven countries were left for analysis. K-Means clustering was performed to find the optimal numbers of country clusters. GDP prediction from other fields was done using different models. Finally, a correlation between all fields was presented as a map where all the fields are compared to each other.

Problem Statement and Hypothesis

The CIA fact book contains information that will be used to predict GDP. This GDP prediction is based on low frequency data (i.e. yearly collected) on fields with unknown relationship to GDP. Predicting GDP more accurately is important in finance. Using yearly data on a variety of fields might reveal trends that are not detected by algorithms that try to predict GDP from financial fundamentals exclusively.

Hypothesis: GDP can be predicted from the CIA fact book using fields that are not directly related to GDP. Using 30 fields from the CIA fact book, it is possible to cluster countries better because it is based on a more comprehensive combination of country characteristics.

Data Extraction

The CIA world fact book is available online in HTML and TEXT format. Text files were used to parse the data into a data frame. The text files are poorly formatted across years and fields, and some files required special manipulation because they were not delimited properly in the files. Some countries have long names for which there is no information of when the country name begins or ends. A curated list of countries from 2014 was used as a starting point to read the files. Using a dictionary with these countries as key, every sub string of each line was check to see if it corresponded to an existing country name. This allowed the detection of any country in a line with a name of any number of words.

The data was placed in a data frame with Multi indexing. The columns have three levels of Multi indexing that corresponds to each field's category, name and year. The rows are referenced with two levels of Multi indexing corresponding to country name and geographical region. All information was exported to a CSV file by the script CIA_get_data.py.

The CIA fact book takes into account 261 countries and territories, thirty fields and for the years 2004 to 2014. Each row in the data frame corresponds to a country. If every row that has at least one null value is dropped, there would only be forty one countries left. That is, only forty one countries kept statistics for all fields between the years 2004 and 2014. Countries as important as the United States had missing fields. This suggested that pre-processing of the data was necessary. In other words, imputing missing values from available ones.

Data Pre-processing

There were null values in the data for many different reasons that could be corrected manually. This corrections were done for a few countries only. Making these corrections allowed to match information found on some fields to the appropriate country. For example, some countries changed their names. Cabo Verde was called Cape Verde before 2013. In 2006, Serbia and Montenegro separated.

Special consideration was given to the European Union. The CIA fact book considers the European Union as a country, but it also gives statistics on the countries that form that union. That means that when doing statistical analysis of all the countries, the European Union data is duplicated. For that reason, all data from the states of the European Union was deleted, and the European Union was considered by itself, and missing field values of the EU were filled in by adding the values of its constituent states. Two small states of the European Union, Cyprus and Malta, had missing or not a number (NAN) values on railways. This presented an interesting question, does the NAN values correspond to statistics that are not collected by countries, or NAN actually corresponds to zero? It turns out that for Cyprus and Malta the NAN corresponded to zero. These two small countries used to have railways, but they are not currently in operation. With this data completed manually after an internet search, the European Union had data on all its fields.

Yet, many important countries had missing data that cannot be easily filled by checking the facts on the internet. A method to fill this missing information using linear regressions was implemented. For countries that have data on at least two of the eleven years in consideration, the missing data was filled from a linear regression from the existing data points. These linear regressions were also calculated for countries that had no missing data.

In order to test the accuracy of the linear regression in predicting values for missing fields. Another linear regression was done on countries that had at least two values from 2004 to 2013. Of these countries, a subset of countries that have statistics for 2014 was selected. In this way, the 2014 data was used as a target to measure the accuracy of predictions done with linear regressions from 2004 to 2013. Figure 1 presents root mean square errors (RMSE) for the prediction on every field. RMSE give a estimate of the accuracy of the statistics. The fields with the highest RMSE are found in the Economy category and correspond to GDP real growth rate, industrial production growth rate and current account balance.

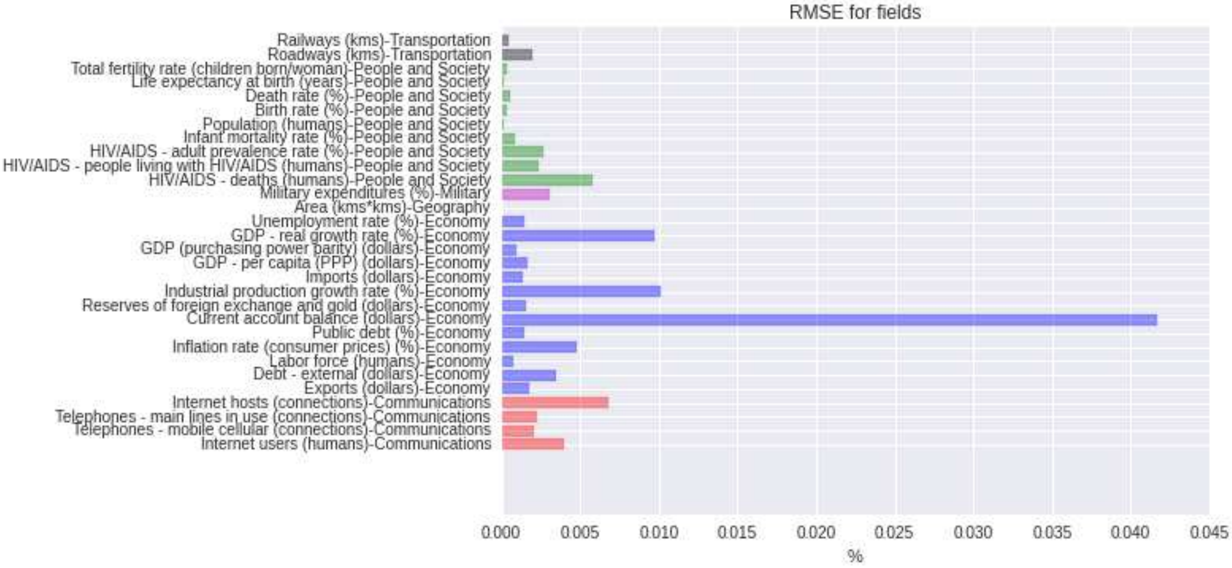

Figure 1: Root Mean Square Errors for each field.

After linear regressions were used to impute, or fill in, missing values, the number of countries that have numerical values in all fields for all years went up to seventy seven. Figure 2 shows the number of NANs or null values for every field between 2004 and 2014. The plot is a heat map, and it color codes the number of NANs for a quick comparison of how the collection of data on some fields in the world varies. Linear regression were done by the script CIA_imputation.py, and exported to a CSV file.

The two final steps in pre processing the data were the regularization and normalization for the fields' values. Data in the fields range across several orders of magnitude, and have different units. GDP is counted in the trillions of dollars, HIV/AIDS statistics are in the thousands of people. In order to compare fields that are measured in different units, we used Unit-based normalization (Equation 1) to make all values positive (from 0 to 1) for each column-field(X). This type of normalization takes care of some fields, such as inflation and GDP growth rate, that can be negative, and it conserves the orders of magnitude of the original data while keeping it a positive number. The next step is to divide the date by the sum of all values in the field to get a percentage of how much a given country compares to the hole country (Equation 2).

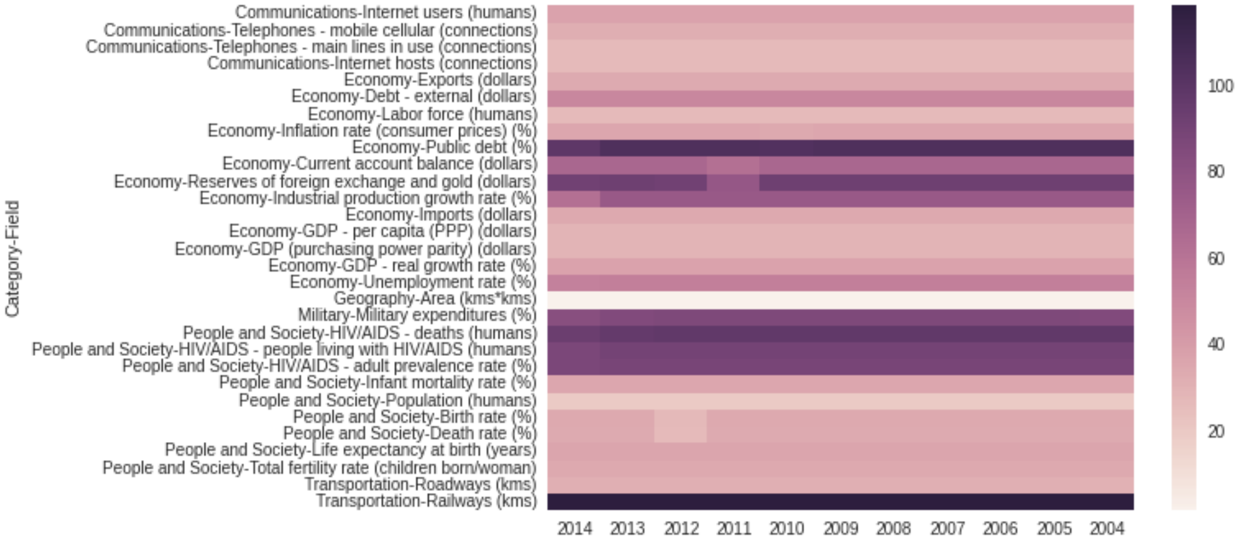

Figure 2: Heat map with number of nulls per year per field.

$$
\begin{aligned}
Z_i &= X_i - \min(X)/\max(X) - \min(X) \quad (1) \\
N_i &= Z_i/\sum X \quad (2)
\end{aligned}
$$

The above two equations transform the data for interpretations. All columns are re scaled so that their values range from 0 to 1 and add up to 1. In this way data can be interpreted as percentages. The following is an example of how to interpret the data across fields: The United States GDP (PPP) is 0.206, or 20.6% of the seventy seven countries' GDP, and it has 60.21% of internet hosts, and 0.823% of the total fertility rate of the world.

Some field/columns were multiplied by -1 to reflect that is not something good for the well-being of people. GDP is positive, HIV deaths,negative; Exports, positive; imports, negative. This is crucial so that when we compare countries in a 30 field space, the euclidean distance should give a measure of difference in progress between countries, and fields that are a bad statistic for a country can not be considered something positive. The term negation is coined for fields that are multiplied by -1 because they are not good for the people of a territory. Normalization, regularization and negation were calculated by CIA_red_n_norm.py.

Clustering

Countries' mean values for every field between 2004 and 2014 were calculated. A dataframe with seventy seven rows and 30 columns (Figure 3) was used as input for KMeans clustering. The silhouette coefficient and the 'within sum of square errors' was calculated

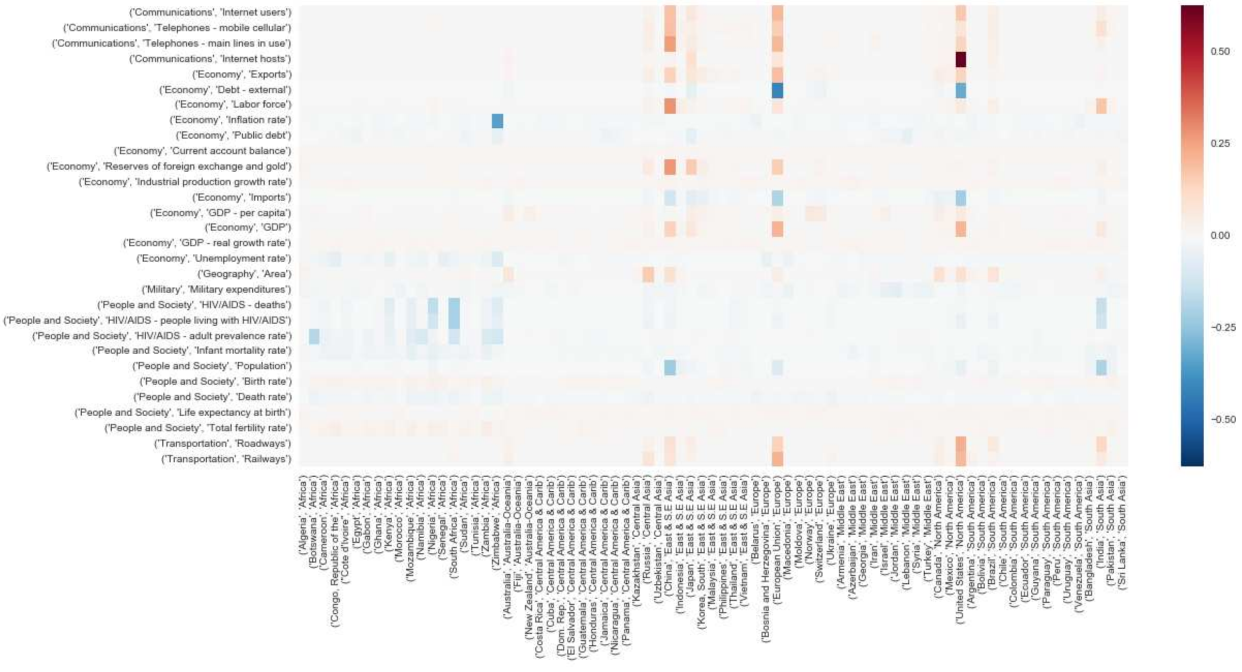

Figure 3: Heat map of all countries and fields after regularization and normalization. Values range from -1 (blue) to 1 (red)

for up to 35 clusters. Figure 4 below shows that 12 clusters is the smallest number of clusters for which the 'within sum of squares' changed significantly. A visual representation of clustering countries in 12 groups is presented in Figure 5. Countries in the same cluster were highlighted with the same color. This division of the world's map is a model based on the classification of the seventy seven countries by thirty different fields. This is a more complex classification than the usual classification by geagraphical proximity.

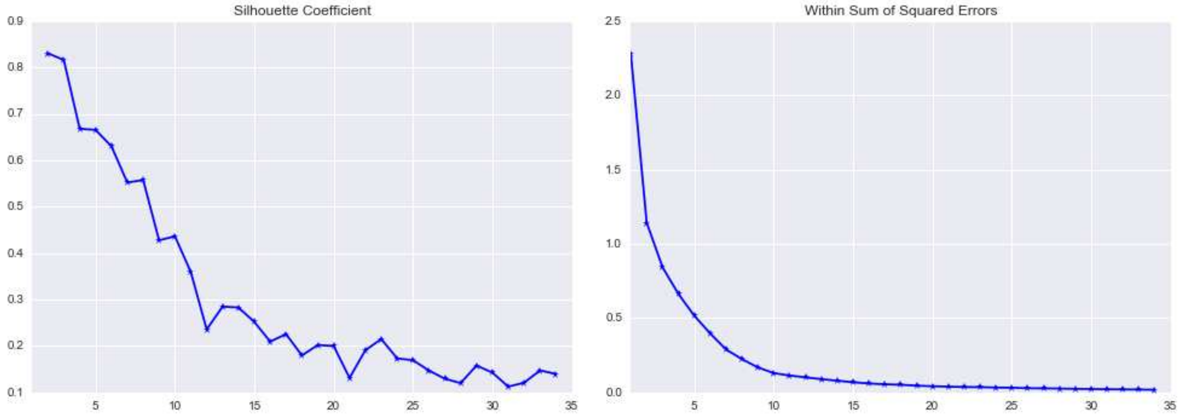

Figure 4: Results for 35 clusters. Left, Silhouette Coefficient. Right, within sum of square errors.

A map of the world illustrating the classification of the seventy seven countries considered is given in Figure 5 below. The most notable feature is that United States (purple), Europe (red), China (light blue), India (pink) and Japan (black) belong to a cluster of their own. This classification is likely due to GDP related fields that carry the most weight. Canada, Australia, Russia and Brazil (brown) form a cluster. These countries have large economies that are significant but trail the super powers that belong to clusters of their own. Zimbabwe is in a cluster by itself (light pink), and this is due to large inflation. Sub Sahara countries such as Sudan, Congo, Cameron (green) form a clusters likely becausethey have high HIV/AIDS rates.

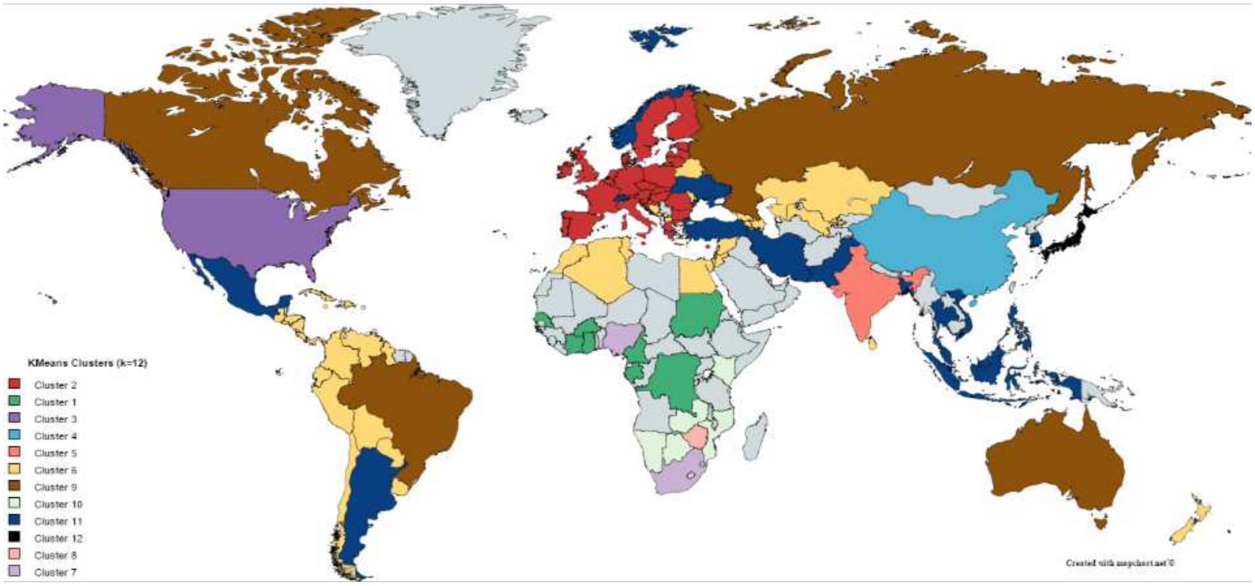

Figure 5: Illustration of best results from K-Means clustering. The number of clusters is 12. Only countries colored in light grey are not part of a cluster, they correspond to countries not considered by K-Means because they have null values that could not be imputed.

GDP Prediction

GDP had already been predicted during imputation of missing data. That prediction was done with 2014 as target value, and data between 2004 and 2013 was used to draw the linear regression and predict GDP on 2014. This was done on a per country basis.

In this section, GDP will be predicted from fields. There are three fields that are a direct measure of GDP; these fields are removed because we want to predict GDP from unrelated fields. The remaining twenty seven fields were used as predictors. The target to be predicted is a variation of GDP growth rate. The field of GDP growth rate was divided in two. Fifty percent of the countries with the highest GDP growth rate were given a value of one, the other fifty percent were given a value of zero. In this way, GDP growth rate gets digitized, and we will use the other twenty seven fields to predict if a country is in the top fifty percent of GDP growth rate or the bottom.

Decision trees, logistic regression and knn together with cross validation and grid search were used to predict GDP. The following table shows the results from those predictions.

Table 1: Accuracy of different models in predicting GDP.

Predictive Model	Accuracy
Logistic Regression	0.553
Knn (n=5)	0.521
Knn (n-12)	0.511
Decision Tree (depth = 2)	0.663
Decision Tree (depth = 3)	0.711
Decision Tree (depth = 10)	0.611
Decision Tree (depth = 3, entropy)	0.765

Decision trees outperform knn and logistic regression for all depths. A grid search was done for up to a depth of 20 (Figure 6), and a depth of three gave the best GDP prediction of 0.711. Yet, it was when using the decision tree with the entropy option and a depth of 3 that the best result was obtained, 0.765.

The prediction of GDP was done on an arbitrary division of the countries in two, defined by a separation according to countries in the top and bottom 50% by GDP. For this reason, a confusion matrix did not make sense. Defining true-positives, false-positives, true-negatives and false negative on an arbitrary division of the countries is difficult to interpret. Instead, the features that are more important in predicting GDP were plotted to find out which fields contribute the most to predicting GDP.

The features that contribute the most to predicting GDP are shown in Figure 7. These features are life expectancy at birth, industrial production growth rate, internet hosts and exports. These features are associated to countries that are important economies in the global scale. China has one of the largest industrial production and exports. Germany also is a very strong exporter. United States is the country with the most internet hosts. Developed countries with significant GDPs have higher living standards and higher life expectancy. This is all related to GDP, and these features that are more important are very revealing, yet not surprising.

Correlation

The best way to represent the relationship between every field and every other field is by plotting a correlation matrix. This gives a more comprehensive view of the relationships among fields, Figure 8. Fields that correlate are represented by a red color; fields that do not correlate, blue.

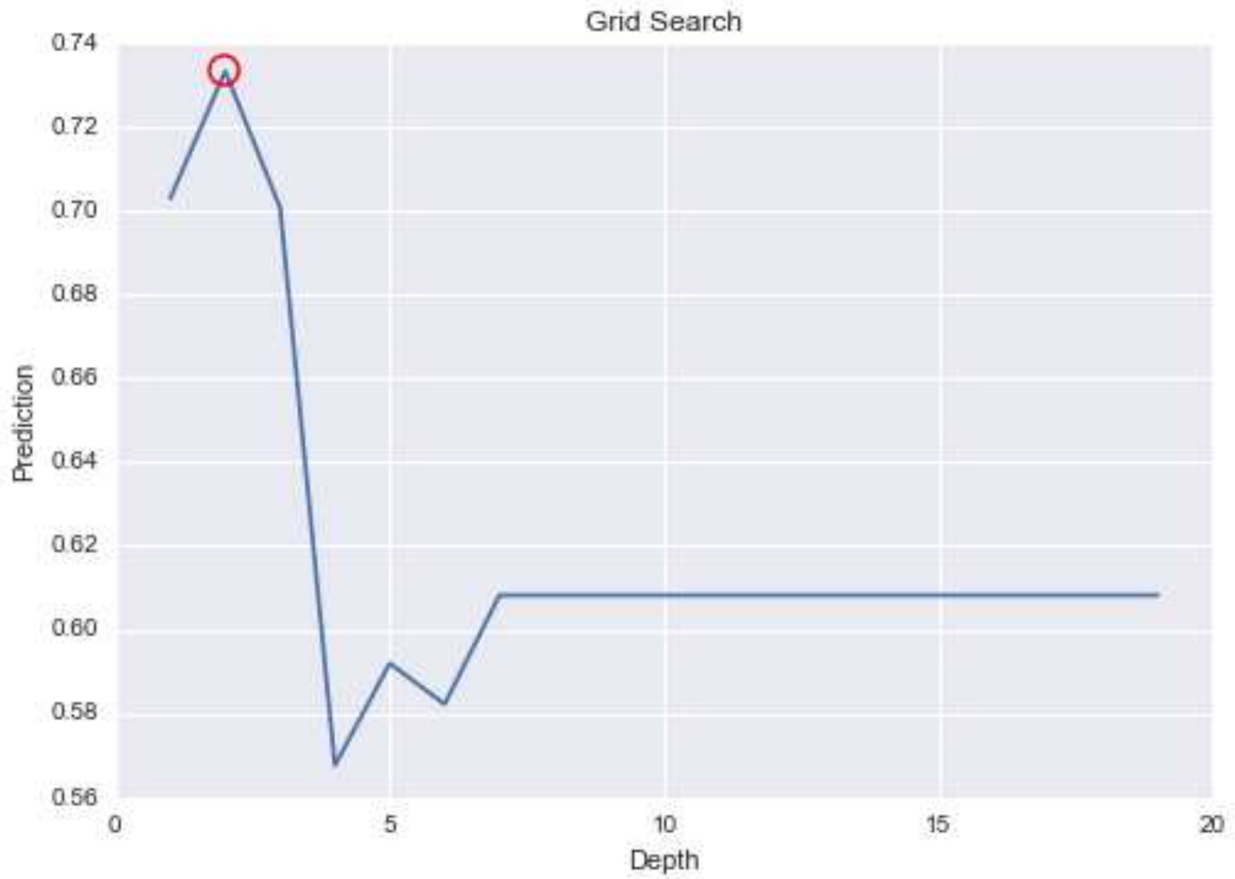

Figure 6: Grid search for decision Trees of different depth. A depth of two is the ideal result for decision trees that do not use GINI.

Conclusions

Lessons from exploring and visualizing the data: Even an intelligence agency that should collect accurate information gathers it in a sloppy manner. This is probably because the data is collected manually and it is laborious. Also, the data is presented as a guide for people, and it is not intended for data analysis. People just ignore null values, programs do not know how to handle them. One lesson is not to assume that a null value is a zero value, or lack of information; it could be either one.

How you chose which features to use in your analysis, and details of modeling process: I wanted to find ways to compare countries that would give a better idea of what differentiate countries with good and bad living standards. The motivation is to find ways to understand the determinant that make countries rich or poor in order to find ways to help their people. Linear regressions was used to impute missing values of fields in order to have a large enough set of countries for comparison. This is also the motivation for GDP prediction, and finding the features that are more important in this

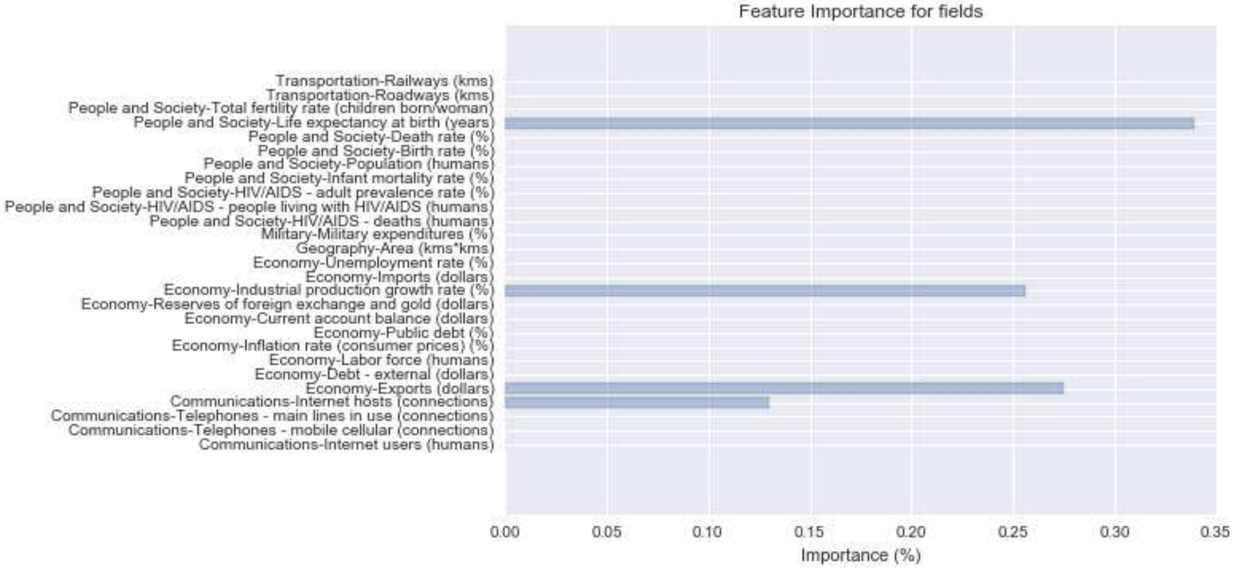

Figure 7: Features, or fields, that are more important in predicting GDP using a decision tree of depth three and gini option.

predictions.

Challenges and successes: Cleaning the data from the CIA website was a big challenge as well as understanding what is in the data. Once the set of countries for the analysis was cleaned of null values, the use of algorithms to predict GDP was relatively easy. Another challenge was using a data frame with Multi indexing. This way of indexing data is more complex than the single indexing level used in the scripts. Multi Index was necessary because there are fields with the same category, and every field has a value for every year between 2004 and 2014. Finally, a challenge and a success was to make the project appealing and simple at the same time. I wanted people to be able to follow what I did, and also to present interesting classification and predictive algorithms on contries' information.

Possible extensions or business applications of your project: I want to put this data on the web for any data scientist to download without having to go to the trouble of cleaning and filling up missing values. I would like to make the website like a wiki so that people can fill in missing values. People would also be able to add new fields to categorize countries, or create new lists of things to rank, such as soccer teams, movies, books or people. It would be more structured than a regular wiki, mostly numbers, or less wordy, and more readily applicable to data analysis.

Conclusions and key learning: The key learning is to be able to take a data set, clean it, analyze it and draw conclusions. This was all done in a rational and structural way, so that we can make sense of the data in the real world, and not just in a statistical realm. The hypothesis that GDP can be predicted from the CIA fact book turned out to be true. GDP was predicted with a 78% accuracy when divided in two groups.

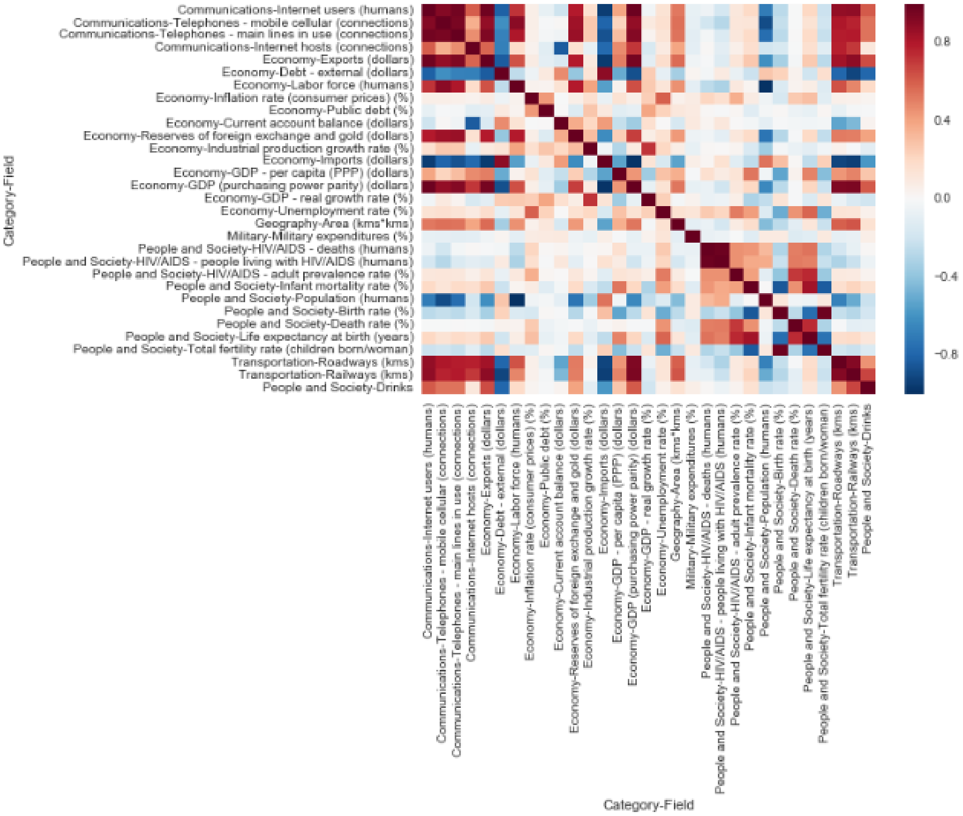

Figure 8: Correlation matrix between all fields. The right most column is not one of the thirty original fields. It is a measure of alcohol consumption.