Data-Oriented Programming Paradigms: Exercise 3

Extreme Poverty Analysis

Alkojjeh Amer 01646631 | Antolín Rodríguez José Carlos 01127194 | Arfaoui Ghaith 01435404 | Shehu Jeton 11902020

INTRODUCTION

The goal of this assignment is to perform an exploratory data analysis and modeling in order to answer the following questions:

- What percentage of the world population lives in extreme poverty?
- Which characteristics are predictive for countries with large populations living in extreme poverty?
- Which characteristics are predictive for populations emerging from extreme poverty?

1. Data Gathering

Multiple datasets are downloaded using the World Bank Api. These datasets represent different indicators about education, gender, health and development for all countries around the world. The different datasets are merged to form a new one and only data between 1990 and 2015 are considered. The merged dataset has 4920 instances and 5476 attributes. For the population data, all information about rural areas and cities that are present in the dataset are removed since we conduct an analysis on a national level.

As ground truth we consider the poverty head count in each country which is calculated based on a daily income of $1.9 and adjusted to the 2011 purchasing power parities (PPP). These values are only available every 2 or 3 years.

2. Missing Data

It was observed that around 75 % of the data is missing. According to the World Bank, most of the data are missing due to conflict, lack of statistical resources, etc. This allow us to conclude that the data is not missing at random.

In order to deal with such huge amount of missing data, we use the forward fill from pandas which propagate last valid observation forward to the next valid one and then the backward fill which uses the next valid observation to fill a gap. This imputation is done country-wise. In addition, all attributes having more than 20% missing data are removed.

3. Correlation Handling

Another issue is the existence of multiple attributes that are highly correlated. To deal with this, we start by removing all attributes that are directly related to the target attribute (head count). Then, we define a function that provides a list with all features having a pairwise correlation higher than a given threshold with respect to a previous feature (index-wise) in the data which is not already in the list. The correlation "pruning" is performed at 0.95 with “Pearson correlation”. The attributes present in the resulted list are then removed.

4. World Population Living in Extreme Poverty

Using the head count and population values, we calculate the population living in extreme poverty in each region. A visualization of evolution of the population living in extreme poverty between 1980 and 2015 can be seen in Figure 1.

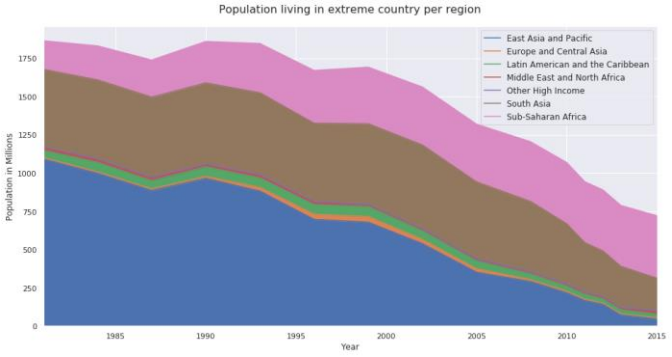

Figure 1: Population living in extreme poverty per region around the world

For the data of 2015, classifying the countries based on the head count level into 8 groups shows that around half of the global population leaving under extreme poverty line live in countries where, on average, at least two thirds of the population live also under the 1.9 extreme poverty line. In 20% of the cases they are even the overwhelming majority where they live (there are 11 countries with around 70% of the population living under the extreme poverty). Most of the rest of extremely poor population (around 41%) live in countries with significant headcounts, from 5% to 30%. Only 5% of the world extremely poor population lives in countries where they amount to 5% or less.

Looking at the data of 1981, back then around 90% of the extremely poor population was concentrated in only 35 countries with an average 74% of population living below the threshold. We can also see that there are significantly more countries with very low levels of extreme poverty.

Looking at the headcount with respect to Gini distribution, we conclude that there is no direct relationship between inequality and extreme poverty. Although, it is more likely for countries with high inequality to have more poverty. Another important fact is that for

20 countries representing 45% of the world poor population there are no Gini measures.

5. Characteristics of Countries with Large Population Living in Extreme Poverty

In order to determine the best attributes for predicting countries with high levels of extreme poverty, we use two methods: Random Forest and selection based on the ANOVA F-value (F-Classif). We observed that both methods provided different results. We then trained a Random Forest classifier using the feature resulted from each method and calculate the explained variance score for each. Results are summarized in Table 1.

Table 1: Variance score for both feature selection methods

Feature Selection Method	Explained Variance Score
Random Forest feature importance	0.729
F-Classif	0.950

Performing hyper-parameter tuning with randomized search resulted in a slight improvement of the explained variance score to reach a value of 0.954.

Therefore, the top 10 attributes that characterizes extremely poor countries are the following:

- Median: the median of monthly household per capita income or consumption expenditure
- Mean: the average monthly household per capita income or consumption expenditure
- Human Capital Index (HCI)
- Specialist surgical workforce
- Adjusted net enrollment rate, upper secondary
- People using at least basic drinking water services
- Risk of impoverishing expenditure for surgical care
- Probability of Survival to Age 5, Female
- Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)
- Agriculture, forestry, and fishing, value added (% of GDP)

For further investigations, countries are classified into 4 categories based on the income level: low, lower middle, upper middle and high income. First, a model is trained considering lower and lower middle income only. The resulted variance score was 0.94. Using the upper middle- and high-income countries only gives a variance of -0.04 and it turns out that there is an outlier. After removing the outlier, the variance score improved to 0.695.

6. Characteristics of Countries emerging from Extreme Poverty

Countries are classified into 5 categories based on the change of the extreme poverty headcount between 1990 and 2015. We calculate the differences in all attributes between the values in 1990 and 2015 for all countries. A country is defined as poor if more than 10% of its population is living in extreme poverty. Categories are:

- Rich Countries (R): Countries that have a value of extreme poverty headcount below 10% in 1990 and in 2015.
- **Emerging Countries 1 (E_R1):** Countries that have a value of extreme poverty headcount above 10% in 1990 and below 10% in 2015 with a decrease in this value of 50% or more (head count in 2015 is half or less than what it was in 1990).
- **Emerging Countries 2 (E_R2):** Countries that have a value of extreme poverty headcount above 10% in 1990 with a decrease in this value of 50% or more (head count in 2015 is half or less than half of what it was in 1990), but with extreme poverty headcount still above 10% in 2015.
- **Non-emerging countries (NE):** Countries where the head count in 1990 was more than 10% and still more than 10% in 2015 with less than 50 % reduction in the head count value.
-Countries going into extreme poverty (B): Countries with less than 10% head count in 1990 and more than 10% in 2015.

This classification can be seen in Figure 2.

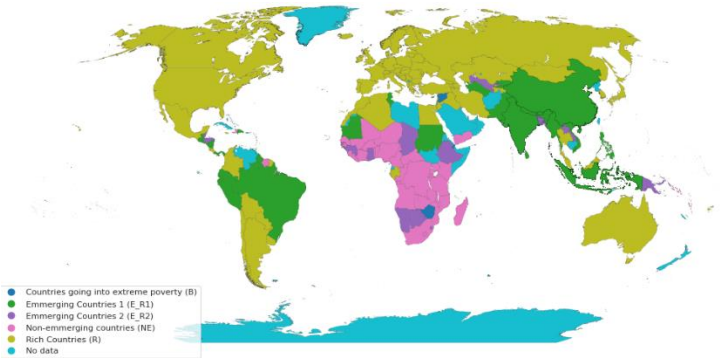

Figure 2: Classification of each country around the world

Once classes are defined, several methods are used to select the attributes that predictive to emerging countries. These methods include Random Forest, LightGBM and Logistic Regression. Since all methods resulted in slightly different attributes, classification models are built based on the features resulted from each method and each performance measures are calculated and results are shown in Table 2.

Table 2: Performance measures for each selection method

Selection Method	Accuracy	Balanced Accuracy	F1
Random Forest	0.810	0.593	0.784
LightGBM	0.707	0.508	0.703
Logistic Regression	0.779	0.588	0.766

It can be seen that the feature selection using random forest delivered the best results. The selected features are:

- Mean: the average monthly household per capita income or consumption expenditure
- Median: the median of monthly household per capita income or consumption expenditure
- Population growth (annual %)
- GDP per capita (constant 2010 US$)
- GDP Per Capita, PPP (Constant 2011 International $)
- Individuals using the Internet (% of population)
- Out-of-pocket expenditure per capita
- Cause of death, by non-communicable diseases (% of total)
- Population ages 25-29, female (% of female population)
- Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)

CONCLUSION

It was observed from 5. And 6. that economical and health related attributes are the best for explaining extreme poverty with the mean and median monthly household per capita income being the most important ones.

REFERENCES

[1] Ferreira, F. H. G. et al. (2016): A global count of the extreme poor in 2012: data issues, methodology and initial results. J Econ Inequal 14:141–172. doi:10.1007/s10888-016-9326-6.
[2] United Nations (1995): Report of the World Summit for Social Development, 6–12.
[3] WorldBank. https://databank.worldbank.org/source/education-statistics
[4] WorldBank. https://databank.worldbank.org/source/gender-statistics
[5]WorldBank.https://databank.worldbank.org/source/health-nutrition-and population-statistics
[6] WorldBank. https://databank.worldbank.org/source/world-development-indicators
[7] WorldBank. http://iresearch.worldbank.org/PovcalNet/povDuplicateWB.aspx
[8] WorldBank. http://iresearch.worldbank.org/PovcalNet/WhatIsNew.aspx