Poverty Level Prediction for Households

1st Chiranth Jawahar
Computer Science Engineering
PES University
Bangalore, India
chiranthjawahar@gmail.com

2nd Namrata R
Computer Science Engineering
PES University
Bangalore, India
namrata.ajjampur@gmail.com

3rd Diya Chandra
Computer Science and Engineering
PES University
Bangalore, India
diyasateesh96@gmail.com

***Abstract*—This project focuses on prediction of poverty levels of households based on socio-economic factors and classifies them into 4 distinct categories. Features selection and level classification on a data set containing household characteristics of a representative sample of Costa Rican households allows accurate insights into features to enable local development in the required areas and appropriate allotment of funds for social welfare benefits.**

***Index Terms*—classification, poverty, random forest, xgboosting**

I. Introduction

Poverty levels have been hard to pinpoint in developing nations due to the scarce availability of data. This problem also occurs due to the lack of any substantial model which accurately represents these categories thereby causing unscientific norms to be used for present day classification leading to a certain amount of misclassification of social welfare benefits. This classification of poverty levels helps governments and organizations allocate social welfare benefits in an efficient and effective manner thereby expediting the process of socio-economic upliftment in the region causing an adverse positive effect in these nations. The classification that takes place also allows clearer understanding of the major pain points of each poverty class by observing the most notable features, this allows for more care to be taken in the handling of those specific factors thereby helping in alliviating indviduals of that status.

II. Previous Work on Similar Problems

A. Random Forest Implementation

Random forests are an ensemble of decision trees where they are trained using a sample of the data and the class that receives the majority of votes is obtained. Random forests run considerably fast and are known to have high accuracy when compared to other classification methodologies. It is taken into consideration that generally random forests do not overfit as for a large number of trees, the generalized error value converges to a limited value under a strong law of large numbers [1].
The method used is in previous work[2] is as follows:

- First a random sample of observations are taken and the subsequent bootstrap samples for other trees are taken.
- A subset of variables much smaller than the total number of variables are taken and using the Gini score the best split is found.
- An out of bag (OOB) prediction is obtained from the majority vote across tree samples whose observations were not taken in the bootstrap sample.

This method relied heavily on the income of individuals as its primary differentiator and hence removed all individuals who did not have any income mentioned. To evaluate the importance of a feature, [1, 3] a technique was proposed by which, for all trees in the forest the average of an impurity decrease measure for all nodes in which the feature is concerned is taken into consideration. The feature which has the largest decrease of this measure is considered to be the most important. They achieve this using the Mean Decrease Gina (MDG).
Cases in which any sort of missing values are present are also removed from the sample. The results obtained show factors that strongly effect whether an individual is poor or not and helps classify them into 2 base categories.There is also a noticeable limitation where a lot of features were completely removed due to the presence of a large number of missing values.
a few issues that can be noticed is that due to multiple cases and features being removed an accurate representation of the population is not truly seen and this may cause the result to report incorrect conclusions. Another case where this methodology has issues is with regard to its static barrier of stating that individuals are only classified into 2 groups which makes it harder to notice and differentiate between extreme cases in the same class which in actuality may need to be treated differently.

B. XGBoosting

XGBoosting is used for classification and detection of epilepsy[4] in a binary format (healthy; patient has epilepsy).This is done by finding the combination of features that show the best predictive power in the binary classification. Xgboosting is an implementation of gradient boosting decision trees. the method for implementation is as follows:

- feature selection is done using filter and wrapper methods. Filter methods allow fast computation and provide feature ranking which facilitates removal of unnecessary features. Wrapper approaches uses a classification algo-

rithm for an evaluation of a subset of features by training and testing with cross validation on the subsets.
- A learning rate of 0.01 for better generalization is taken with xgboost[5].The number of boosting trees that are used as estimators is set to 1200. A subsample value of 0.7 over the default value of 1 is taken to prevent overfitting. The model's complexity is also decreased by making maximum depth 3 over the default depth of 6.

The results that this model received were fairly positive with the AUC being the performance metric with a value of 0.91. A few notable issues were with present as feature selection in their case was moderately unsuccessful, making the classification with the initially employed methods inaccurate, due to this further cross validation was required. This scenario also consists of almost perfect data which required little to no preprocessing, which is not the case in most real world scenarios such as the problem being addressed.

III. Proposed Problem Statement

The need for classification of households based on their situational factors into specific levels of poverty (or absence of it) to help in allocation of social welfare benefits suitably amongst the population and to help identify the factors that strongly affect different levels for general upliftment.

IV. Problem Approach

The data of each individual in a household is acted upon along with their collective features. Poverty levels are broadly categorised into 4 tiers which are,

- Not vulnerable
- Mildly vulnerable
- Vulnerable
- Extreme poverty

The initial idea is to implement xgboost on the preprocessed data with neural networks to provide an ensemble allowing for greater accuracy. The feature selection process is taken care of by looking at filtering, wrapping and embedded methods to provide an accurate ranking and weightage of each of the features allowing for removal of unnecessary features, the idea of going with principle component analysis is also being explored.

A. Data

The dataset used is from Kaggle which is an online community of data scientists and machine learners. The training dataset contains 143 columns and 9557 rows. Each column in the dataset represents the various features that most likely affect the households. Each row represents a member of a given household and the information of the various features.The target attribute has been classified in to classes indication their poverty status. Each household is uniquely identified by the attribute "idhogar" and each row is uniquely identified by an "Id". The columns indicate whether a household has access to certain amenities, provides insight to their educational background and their present living conditions. The squared values of certain features are also provided as separate columns.

B. Data Cleaning and Preprocessing

- The dataset is grouped based on the household that the members belong to and each member of the household is reassigned the same "Target" class as the head of the household.
- Handling attributes with non-numeric values- dependency(dependency rate), edfeje(years of education of male head of household) and edjefa(years of education of female head of household). The values "yes" and "no" are mapped to integer values 1 and 0 respectively for the above mentioned columns. "Id" and "idhogar" are not altered as they uniquely identify the rows.
- Analysing and handling missing values.
  For the column "rez_esc"(Years behind in school), members who do not belong to the age group of 7-18 are assumed to not be attending school and hence a 0 value is set to the missing values. For those within the given age group, the value is replaced by the mean of the those already present in the dataset.
  For the column "v18q1" (number of tablets owned by a family), the missing attributes are filled with 0 as they indicate that no tablets are owned by the family, which can be verified by comparing with "v18q" column(owns a tablet or not)
  For the column "v2a1"(Monthly rent payment), The missing values are are replaced by 0 if the head of the household owns the house.In other cases, the mean rent payment for each group of poverty level is found and missing values are replaced based on the class of poverty that the household belongs to.
  For the column "meaneduc"(average years of education for adults (18+)), "meaneduc" is calculated by grouping the members based on household, finding the mean of escolari(no of years of education) of all the members whose age is above 18. Similarly the "SQBmeaned" column is also filled with the sqaured value of the previously discussed attribute "meaneduc"
- Attributes with duplicate values are listed : "hhsize" "hogar_total" and "agesq". These attributes are dropped from the dataset

Finally, two dataframes are created for further analysis, EDA and statistical modelling. one with the attributes are discussed above and another, where attributes with values corresponding to the square of certain other attributes are dropped.

References

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[2] Ruben Thoplan, "Random Forests for Poverty Classification" in International Journal of Sciences: Basic and Applied Research (IJSBAR), August 2014

[3] L. Breiman, "Manual on Setting Up, Using, And Understanding Random Forests V3.1," Technical Report, 2002.

[4] L. Torlay, M. Perrone-Bertolotti, E. Thomas, M. Baciu, Machine learning–XGBoost analysis of language networks to classify patients with epilepsy

[5] Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367–378