Costa Rican Household Poverty Level Prediction

Scott Zhao
miz107@ucsd.edu

Yuchun Hsieh
yuh378@ucsd.edu

Qihan zhang
qiz237@ucsd.edu

Gautam Nain
gnain@ucsd.edu

1. Introduction

The main objective of this problem is to predict the Poverty level of a costa rican household based on the socio-economic parameters provided in the dataset. This challenge is part of data science for good machine learning competition on Kaggle. Its a supervised machine learning task. The poverty labels are divided into four levels making this a supervised multi class classification problem. The main goal of this problem is to improve traditional methods to identify families at need of aid. So, before proceeding with the machine learning task we can analyze the dataset with various plots and charts to find the relation between the various parameters and the target labels. For example, how education relates to poverty level and how various other economic factors like having a television set in a household help us in determining the poverty level. We can use matplotlib and seaborn python libraries to get better insights into the dataset.

2. Related Work

There has been many submissions for this problem on Kaggle and we would try to solve the problem in a better way than the previous submissions.

3. Dataset

- IDB Costa Rican Household Dataset: This dataset was provided by Inter American development bank as part of the challenge on Kaggle. It provides us with the different socio-economic indicators of 9558 Households to train the model to classify their Poverty level. In the test dataset they have socio-economic data for 23K Households.

4. Project Breakdown

This Project can be broken down into four main components as mentioned below:

4.1. Pre-Processing Data/ Data Cleaning

Before doing data analysis, our first goal is to import the data to python and pre-process it, so that it can be used to be for data analysis and model training. we need to understand the data better and bring it to a desired format that can be used to train the model. Pre-processing is also important because there could be many missing elements in the dataset depending on the parameter. We can take help of pandas library to pre process the dataset.

4.2. Data Visualization/Classification Techniques

We need to find good features to solve the classification problem. we can analyze the impact of various indicators individually (univariate) on the target labels or combine them (multi-variate) to analyze their impact on the poverty levels. After selecting the features, we can also analyze the variables using correlation heatmap and parisplot to get the relation between the variables and target labels. The plots and charts help us find important socio-economic parameters to predict the poverty levels of the given households. The graphs also help us get insights on the relation between the social factors that affect an households income levels.

4.3. Model Comparison

After doing the data visualization part we can compare our visual insights with the various machine learning models for this dataset and observe how our data insights stand against standard machine learning techniques like logistic regression and random forest.

4.4. Final Analysis

Our final insights will be based upon our relational data visualization and standard machine learning models. We also need to develop some intuition through our data visualization methods to predict the poverty levels based on the given socio-economic indicators.

5. Project Timeline

- 11th November: Survey related work to analyze how others have tried to solve this problem and what we can do differently to get better insights.

- 14th November: Complete Data Cleaning and preprocessing.

- 21st November: Plot and visualize all the graphs, charts, parisplots and heatmaps.
- 28th November: Display the Insights in an interactive manner and perform comparison of the insights gained through various data visualization techniques and standard machine learning models.
- 5th December: Prepare Final Report.

6. Division of Labor

Each one of us will contribute equally towards the different topics mentioned above. But to finish this project efficiently, we will create sub topics for each of the main topics and divide the work equally. we will get to know more about the sub topics as the project progresses.

7. Tools Used

To implement this Project, we will be coding in Python and using the existing `scikit`, `numpy`, `pandas`, `seaborn` and `matplotlib` libraries.