COSTA RICAN HOUSEHOLD POVERTY LEVEL

Belén Sánchez

Executive Summary

The Inter-American Development Bank is asking the Kaggle community for help with income qualification for some of the world's poorest families[1]. The challenge consists on identifying which households have the highest need for social welfare assistance.

[1] KAGGLE, 2018. "Costa Rican Household Poverty Level Prediction". https://www.kaggle.com/c/costa-rican-household-poverty-prediction

Background

Many governments run social or welfare programs welfare programs that include cash assistance, healthcare and medical provisions, food assistance, housing subsidies, energy and utilities subsidies, education and childcare assistance, etc. When doing this, governments at the federal and state level face a similar challenge. They have a hard time making sure the right people are given enough aid. This is a political sensitive topic since it focuses on the poorest segment of the population, many of whom typically can't provide the necessary income and expense records to prove that they qualify.

In Latin America, one popular method uses an algorithm to verify income qualification. It's called the Proxy Means Test (or PMT). This method considers a family's observable household attributes like the material of their walls and ceiling, or the assets found in the home to classify them and predict their level of need. While this is an improvement, accuracy remains a problem as the region's population grows and poverty declines.

To improve on PMT, the IDB (the largest source of development financing for Latin America and the Caribbean) has turned to the Kaggle community. They believe that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT's performance.

Data

As part of the Kaggle competition the IDB provides with the following data sets.

- train|test.csv - the training set. This is the main table, broken into two files for Train (with a Target column) and Test (without the Target column). One row represents one person in our data sample. Multiple people can be part of a single household. Only predictions for heads of household are scored.
- sample_submission.csv - a sample submission file in the correct format. This file contains all test IDs and a default value.

Update

1. EDA started.
    a. Explored features with basic statistics.
    b. Identified null values.
    c. Cleaning process of null values started.
2. Models. So far, no models have been run.

Timeline

During the next two weeks, I will follow the next plan:

ACTIVITY	WEEK 1 (OCT 8 – OCT 12)	WEEK 2 (OCT 15 – OCT 19)	WEEK 3 (OCT 22)
DATA CLEANING	X		
MODELING		X	
PRESENTATION			X