**Costa Rican Household Poverty Level Prediction**

*Project rationale*

Accurate targeting of vulnerable households is a fundamental component for efficient allocation of humanitarian resources. The objective of vulnerability analysis is to identify who cannot meet their needs and determine their eligibility for inclusion as beneficiaries into social protection programs.

Targeting methodologies vary in complexity from using very simple and direct selection criteria such as age or household composition to involving more intricate techniques for inferring poverty level. Proxy means test (PMT) methodology uses a set of indicators (referred to as proxies) to estimate household welfare in terms of income and/or expenditure[[1]](#footnote-1). Indicators are selected based on modelling of data which often includes demographic- and human capital characteristics as well as productive assets. The final model is often implemented practically with enumerators visiting households to collect information and evaluate it according to agreed weights whereby a composite score is calculated as an indicator of vulnerability. Households that score below a given threshold are found eligible for inclusion into a social protection program.

The quality of a PMT model is very susceptible to decay as changing circumstances may cause data on needs and vulnerabilities to lose relevance and representativeness. This leads to conceptual drift where interpretations of data changes while the overall distribution of the data remains the same. Consequently, the previously modelled decision boundary for program eligibility will start to diverge with respect to new incoming data, eventually resulting in below acceptable predictive ability and the need for a re-modelling of data. Successful and reliable use of PMT methodology not only requires an accurate model, but also a high level of adaptive capacity. For this reason, the use of PMT targeting necessarily moves beyond data modelling to involve the implementation and deployment of a system.

Effective use of PMT methodology for vulnerability targeting requires (i) a highly accurate model, (ii) with the capacity to receive, process and evaluate household information, (iii) a means of communicating results to its users as well as (iv) a system for retaining incoming data and results for continuous performance monitoring.

*Technical implementation*

All data related projects follow the same general process with regards to data extraction, transformation, analysis, and output. This makes the data processing workflow amenable to standardization in terms of its technical organization. The data project template[[2]](#footnote-2) provides a structure which facilitates deployment of a PMT model as a service with which users can communicate. It currently includes a standardized project structure with a clearly defined execution process and logging functionality. As such it provides a framework within which the project specific details of data management (data ingestion, cleaning, and analysis) can easily be included.

With this framework in place, the next step in the PMT system implementation is data cleaning and validation according to the Inter-American Development Bank data description and requirements.

**Data description and validity requirements:**

The original dataset contains a total of 9557 rows, with 143 columns and consists of information on individuals, where each row represents one person in the sample. Multiple people can be part of a single household, but the level of poverty *will only be predicted for the heads of household*. This structure is represented by the following core data fields:

Table 1: Core data fields

|  |  |
| --- | --- |
| **Id** | Unique row identifier (individual identifier) |
| **Idhogar** | Unique household identifier (group identifier) |
| **Parentesco1** | Boolean indicator of head of household |
| **Target** | Ordinal variable income level:   1. extreme poverty 2. moderate poverty 3. vulnerable 4. non-vulnerable |

**Test requirements:**

* All heads of households must have a poverty level (target); missing values must be discarded.
* All household members must have the same poverty level (target); in diverging cases the poverty level of the head of household (parentesco1 = 1) is normative.
* Each individual must belong to a household (idhogar) which has one head of household; households with no head must be discarded.
* Write a pre/post processing test-suites

**Feature engineering:**

Exploratory analysis may reveal any interesting patterns or trends that may prove useful for feature engineering and modelling.

**Establish baseline model:**

1. https://olc.worldbank.org/sites/default/files/1.pdf [↑](#footnote-ref-1)
2. https://github.com/Cawiess/data\_project\_template [↑](#footnote-ref-2)