Final Project Proposal

DS502/MA543: Statistical Methods for Data Science

Group 3: Claire Danaher, Russell Davis, Renee Sweeney

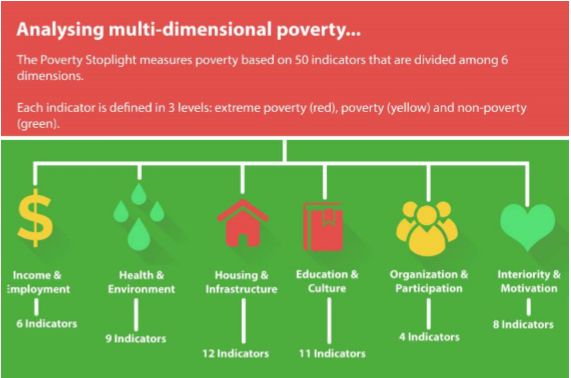
Fundacion Paraguaya is an NGO dedicated to eliminating poverty in Paraguay by addressing problems of unemployment and underemployment. The Stoplight program is an innovative program whereby residents work with program staff to self-assess their poverty based upon 50 indicators, often with subsequent follow-ups to evaluate changes in their situation. Provided below is a summary of the data dimension of the data set.

Table 1: Data Dimension

|  |  |  |
| --- | --- | --- |
| Description | Number of Records | Number of Participants |
| All Data | 12146 | 9486 |
| Data to be Analyzed | 4980 | 2496 |

The data set to be analyzed excludes participants for whom follow-up data is not available. Therefore, our team will be analyzing 4980 records associated with 2496 individuals. The primary variables of interest are the indicators. The figure below provides a summary of the indicators collected by the program.

Figure 1: Multidimensional poverty analysis

[[1]](#footnote-1)

Additional variables include:

* home office
* family income per capita
* urbanization of home area
* date of enrollment in the program
* details about how often they had worked with social workers and loan workers to work on those indicators

The primary purpose of our analysis will be to use the data provided to predict whether or not involvement in the program improved an individual's poverty outlook. This analysis will be done using classification approaches whereby success is defined as an improvement of an individual’s aggregating poverty index score. The aggregated poverty index score ranges from a minimum of 50 (extremely poor for all 50 indicators) to maximum score of 150(not poor on all 50 indicators).

Since our data contains mostly ordinal values, we need to use models that work best with qualitative data in high dimensions. We plan to try multiple logistic regression, random forests, boosting, soft margin classifiers, and support vector machines. Forward and backward stepwise subset selection will be used for dimension reduction as part of the multiple logistic regression model selection process. Additionally, our group conducted a brief literature search to identify other potential approaches used in academia to analyze this type of data. We found a paper about poverty in Pakistan that uses Categorical Principal Component Analysis for dimension reduction[[2]](#footnote-2). Our team plans to try and implement this framework as part of our project.

Classification error and ROC curves will be our two primary error metrics. We will use k-fold cross validation to tune the hyperparameters for the following methods: for random forests, the number of trees and number of predictors available for splitting at tree nodes; for boosting, the shrinkage parameter and tree depth; and for soft-margin classifiers, slack. To determine the appropriate kernel for SVM, we will employ different visualization techniques as well as test out the most popular kernels--linear, polynomial, and radial.

Comments and concerns on this project include that we are still waiting for clarification from the program as to identification of which variables within the data set correspond to which poverty indicators. The program staff has been responsive, so we expect this issue to be resolved shortly. Additionally, we hope to be able to present on the last presentation day as many of the models we intend to use are covered during the last few weeks of classes.

1. Growing with El Major Impact Report, Fundacion Paraguya, August 2016 [↑](#footnote-ref-1)
2. Assessing Poverty with Non-Income Deprivation Indicators: Pakistan, 2008-09, Haroon Jamal. Link: http://www.jstor.org/stable/23617743 [↑](#footnote-ref-2)