Problem Set 2: Predicting Poverty

"Wars of nations are fought to change maps. But wars of poverty are fought to map change" M. Ali

Due Date: Sunday, December 15

1 Introduction

This problem set was inspired by a recent competition hosted by the world bank: Pover-T Tests: Predicting Poverty. The idea is to predict poverty in Colombia. As the competition states, *"measuring poverty is hard, time consuming, and expensive. By building better models, we can run surveys with fewer, more targeted questions that rapidly and cheaply measure the effectiveness of new policies and interventions. The more accurate our models, the more accurately we can target interventions and iterate on policies, maximizing the impact and cost-effectiveness of these strategies."*

The objective is to predict poverty at the household level. Data, however, are provided at the household and individual levels. You can use individual-level information to build extra variables to improve your prediction.

The data comes from DANE and the mission for the "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". The data contains four sets divided into training and testing at the household and individual levels. You can use the variable `id` to merge households with individuals. You will note that some variables are missing in the testing data sets; this is designed to make things a bit more challenging. More information about the data is available at the competition website.

An essential dimension for policymakers is they can *rapidly and cheaply* measure poverty. When building your model, aim for a model that uses the minimum number of variables.

There are two expected outputs:

1. A `.pdf` document.
2. Submissions with your team's predictions in Kaggle at the following link.

1.1 General Instructions

The main objective is to construct a predictive model of household poverty. Note that a household is classified as

$$Poor = I(Inc < Pl) \tag{1}$$

where I is an indicator function that takes one if the family income is below a certain poverty line.

This suggests two ways to go about predicting poverty. First, approach it as a classification problem: predict zeros (no poor), and ones (poor). Second, as an income prediction problem. With the predicted income, you can use the poverty line and get the classification. You will explore both routes in this problem set.

The document must contain the following sections:

- Introduction. The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.
- *Data.*[1]. When writing this section up, you must:
    1. Describe the adequacy of the data to solve the predictive question, the sample construction process, including how the data was cleaned, combined, and how new variables were created.
    2. Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.
- Models and Results. This section presents the specifications and models used for the predictive tasks. Here you must include two subsections:
    1. Models. This subsection describes briefly the models used. The description should include the variables used, how the models were trained, hyper-parameters selection, if any sub-sampling strategy used (and how it was used) to address class imbalances, and any other relevant information. You must submit in Kaggle at least eight (8) predictions trained with at least four (4) of the following algorithms: Linear Regression, Logit, Elastic Net, CARTs, Random Forest, and Boosting.

[1]This section is located here so the reader can understand your work, but probably it should be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here.

2. Evaluation of Results. Here, you describe the results of your models in terms of the relevant metrics. This subsection must include:
   - A discussion with a comparison of the training and test performance of the models. Highlighting the model with the best performance.
   - A detailed description of the model with the best performance submitted to Kaggle.
   - A discussion of the relative importance of the predictors in the best performing model.

- Conclusions and recommendations. In this section, you state the main takeaways of your work.

2 Additional Guidelines

- Predictions have to be submitted on Kaggle. Check the competition website for more information.
- Turn a `.pdf` document to ignaciomsarmiento@gmail.com.
- The document should not be longer than 10 (ten) pages and include, at most, 8 (eight) exhibits (tables and/or figures). Bibliography and exhibits don't count towards the page limit. You are welcome to add an appendix, but the main document must be self-contained. Specifically, a reader should be able to follow the analysis in the paper and be convinced it is correct and coherent from the main text alone, without consulting the appendix.
- The document must include a link to your GitHub Repository.
  - The repository must follow the template.
  - The README should help the reader navigate your repository. A good README helps your project stand out from other projects and is the first file a person sees when they come across your repository. Therefore, this file should be detailed enough to focus on your project and how it does it, but not so long that it loses the reader's attention. For example, Project Awesome has a curated list of interesting READMEs.
  - Include brief instructions to fully replicate the work.
  - The main repository branch should show at least five (5) substantial contributions from each team member.
  - The code has to be:
    * Fully reproducible.
    * Readable and include comments. In coding, like in writing, a good coding style is critical.

- Tables, figures, and writing must be as neat as possible. Label all the variables included. If you have something in your figures or tables, I expect they are addressed in the text. Tables must follow the AER format.