**Exploring the dataset:**

The aim of this investigation is to develop a prediction model that categorizes individual poverty levels into one of four possible ranks.

To better understand the dataset, I read about IDB’s PMT, and how they define poverty. I also read about Costa Rica’s poverty[[1]](#footnote-1) to get a high-level understanding of the socioeconomic landscape we’re delving into with this prediction problem. I also plotted interaction graphs to explore the relationship between select variables that I thought might be important in predicting poverty of a household (e.g. education, dependence, sanitation, roof condition, etc.) to get a better grasp of variable influence on outcome.

It is also worth mentioning that 63% of the training dataset was category-4 (C4), which would likely indicate a high predictive accuracy for true C4 than for a less frequent category such as C1; which is only 8% of the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Total sample | 1 | 2 | 3 | 4 |
| 7646 | 589 | 1296 | 936 | 4825 |
| 100% | 8% | 17% | 12% | 63% |

**Data cleaning and model selection:**

I narrowed down the potentially useful models based on the categorical nature of the predicted variable. The methods I for exploration were boosting trees, random forest, KNN classification, LDA, QDA, multiple logistic regression. I then split the training data into 80/20 train/test proportion.

After graphing interactions between the outcome variable and individual predictors, I selected a few predictors that seemed to have strong influence of the outcome category, then ran the multiple logistic regression formula on those selected variables. The resulting F1 score was 0.305, which turned out to be low once I tried trees as explained later in this report. My approach to fitting the multiple logistic formula has limitations: it relies on my individual judgement, intuition, and level of understanding of the data.

Although I thought KNN classification with CV would be computationally intensive, I ran it to see if it was yielded a good prediction - and because it is easier to interpret to a wide audience. The resulting CV error rate (0.359) was high enough to rule it out as a candidate - this was understandable given KNN’s limitations with high dimensional and mixed data. It also took an unreasonable amount of time to run and find the best KNN (59 neighbours) on a 32Gb RAM computer. Following a similar intuition, I decided to set QDA and LDA aside; and focus on trees and random forest instead.

I then ran 4 iterations of boosting trees using: 1- household data (HH) and out of bag (OOB), 2- (HH) and (CV), 3- individual (I) and (OOB), and 4- (I) and (CV). Within each of those I experimented with number of committees - the higher the number, the better the F1 score, but the longer the computational time (especially with CV which tended to be slightly more accurate than using OOB). The highest F1 I got with boosting was 0.688, which was more than double that of MLR or KNN. I also generated a graph (fig.1) with variable importance to improve my understanding of the features. The top two were mean adult education and squared dependence; while other seemingly important features such owning a computer turned out to be not important at all.

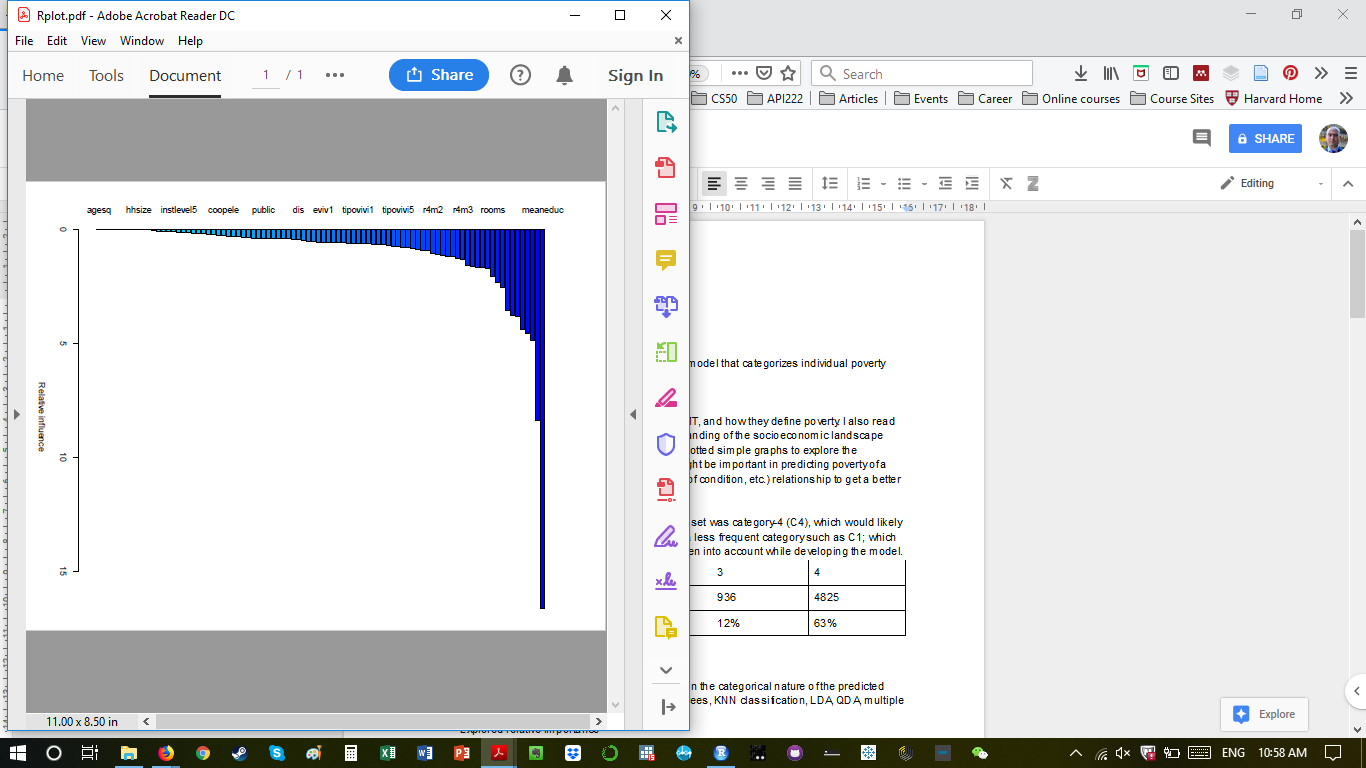


Fig.1 - feature relative importance

I then ran random forest with feature importance. The resulting average F1 score - as well as the individual F1 scores - was high enough to accept random forest as the best model for this prediction problem. Random forest is better than boosting because it decorrelates the variables and uses a smaller number of predictors for each tree. This helps mitigate the dominance of strong predictors in the model, which yields better predictions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| F1-Scores | F1.1 | F1.2 | F1.3 | F1.4 | F1 Avg |
| Random Forest | **0.872** | **0.905** | **0.896** | **0.980** | **0.913** |
| Boosting 2K; CV | **0.644** | **0.668** | **0.558** | **0.884** | **0.688** |

In conclusion, the best method to use in this situation would be random forest, because it yields an average F1 score of about 0.90-0.92.

**Limitations:**

Choosing random forest sacrifices interpretability for gains in predictive accuracy, so it only interprets variable importance; which might not be very useful to explain this to a public audience. Also, because 63% of the data is category-4, nearly all tried models in this paper consistently yielded better F1.C4 scores than F1.C1 and F1.C3 scores (which are less frequent in the dataset). As such, users of this model can be comfortable with its C4 predictive accuracy, but should be cautious with C1 and C3 predictions.

1. Oviedo, Ana Maria, Susana M. Sanchez, Kathy A. Lindert, and J. Humberto Lopez. 2015. Costa Rica’s Development: From Good to Better. Systematic Country Diagnostic. Washington, DC: World Bank. License: Creative Commons Attribution CC BY 3.0 IGO [↑](#footnote-ref-1)