William Bickelmann

Professor Ousley

2/26/18

Research Project

Pover-T Test Competition

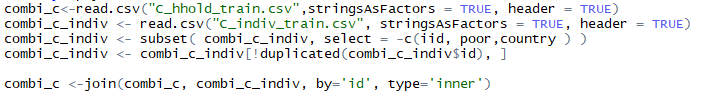
## Introduction

The pover-t test competition is a competition sponsored by the world bank as a way of understanding which strategies work in the reduction of poverty, the world bank invited data scientists to find the most effective way of understanding and predicting poverty based on survey data collected by world bank. The data came from three developing countries A, B and C and each country has dataset organized by individual (each observation is an individual) and by household (each observation is a household). The ultimate task is to predict the probability of poverty of each household in all three countries

## Data Preparation

Part of the challenge of this competition is to conduct the preparation, cleaning and analysis three times. Since the datasets associated with each country has different variables (and also a different number of variables) it is not a simple matter of joining or binding the datasets together and performing one analysis. However, the techniques and processes for cleaning and analysis were nearly identical for each country.

In order to attain the maximum predictive power for the model, the data from both the household and individual data must be utilized. To do this, I used the inner\_join function from the dplyr package to create an inner join on the household ID column and eliminated duplicate household id observations from the individual dataset (since many individuals surveyed were from the same household).



The next step was to eliminate any duplicate columns. Once this was done the id columns was deleted from the resulting data frame.

## Data Cleaning

Cleaning the data primarily consisted of removing NAs and feature engineering. Removing NAs was done in two steps. The first one was to eliminate columns that had over 10 percent of the observations as NAs. After the first step was completed rows with NAs could be removed without much data loss.

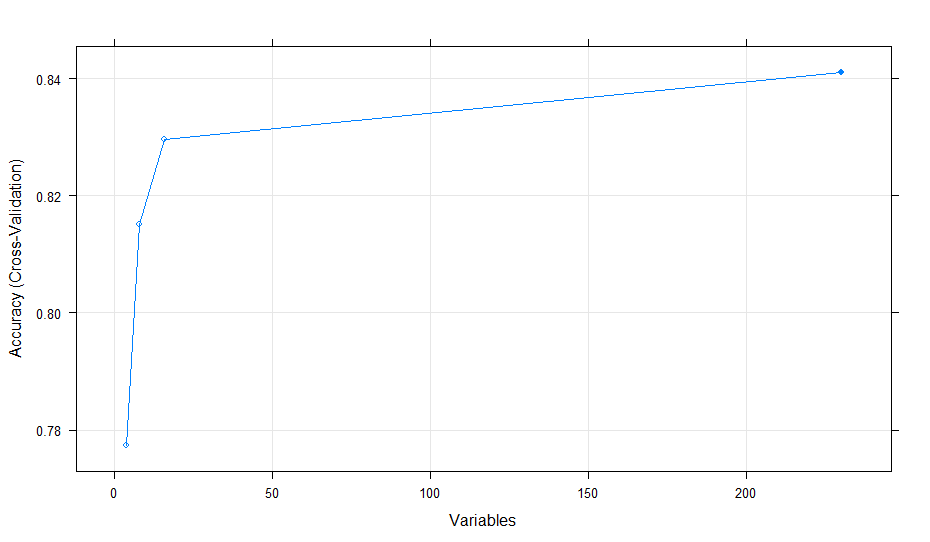
The feature selection took several steps. The function nearZeroVar was helpful to eliminate variables that had little to no variability. The number of variables reduced for each country varied:

Country A from 386 Variables to 230

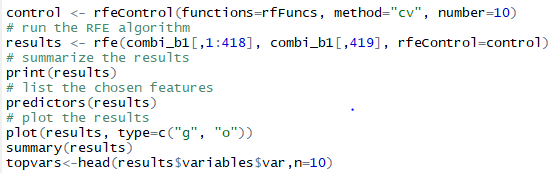
Country B from 665 Variables to 420

Country C from 205 Variables to 177

The next step of feature engineering was the recursive feature elimination which inform which variables contribute the accurate prediction of the predictor variable.

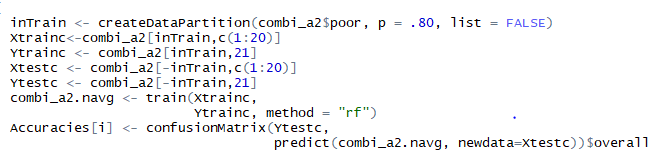


Above is the visualization which demonstrates the number of variables which contributes most to accuracy. Once this is visualized, one can select the number of variables to train the model on.



## Training the Model and Making a Submission

Once the feature selection is complete, one can train a model and make a prediction to submit to Driven Data. The models used included Linear Discriminant Analysis, Flexible Discriminant Analysis, Logistic Regression, and Random Forest. This was done with the following code.



The train function from caret was used to train the model with each method described earlier. Accuracy was used to measure success before using the model which produced the highest accuracy to submit to driven data.

In order to make a submission, household and individual test files needed to be retrieved from driven data for each country. Each test file matched the format of the corresponding train file except that each was missing a ‘poor ‘ variable. The test files were joined using an inner join and the variables which were used to train the model were selected out of the test variables. These variables were used to make a prediction for the probability for each household being poor. This prediction was made into a data frame along with the corresponding household id and country and then submitted to driven data. The best submission yet can be seen below.

