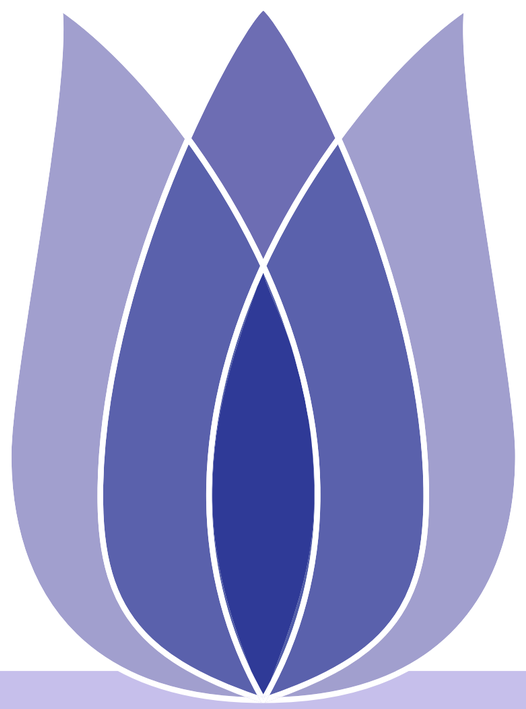
Flip00 Project Final Presentation

Zebin Ju
Xi'an Shiyou University

November 29, 2019



Problem Description

Description

Many social programs today have difficulty ensuring that adequate assistance is provided to the right people. This is especially tricky when a project focuses on the poorest. The poorest people in the world are often unable to provide the necessary income and expenditure records to prove that they are eligible. We need to find an algorithm to predict the poverty level of families.

Problem Analysis

Data Analysis

The data is divided into two files: train.csv and test.csv.

- Key fields:
- ID: an identification of each sample data
- Idhogar: the unique identification of each family
- Parentesco1: identify whether the member is the head of the household
- Target: poverty of families
- 1 for extreme poverty
- 2 for moderate poverty
- 3 for vulnerable households
- 4 for non vulnerable households

Data Glance

	Id	v2a1	hacdor	rooms	hacapo	...	SQBovercrowding	SQBdependency	SQBmeaned	agesq	Target
0	ID_279628684	190000.0	0	3	0	...	1.000000	0.0	100.0	1849	4
1	ID_f29eb3ddd	135000.0	0	4	0	...	1.000000	64.0	144.0	4489	4
2	ID_68de51c94	NaN	0	8	0	...	0.250000	64.0	121.0	8464	4
3	ID_d671db89c	180000.0	0	5	0	...	1.777778	1.0	121.0	289	4
4	ID_d56d6f5f5	180000.0	0	5	0	...	1.777778	1.0	121.0	1369	4

5 rows × 143 columns

Figure 1: data characteristics table

Explore Tags

Target is the tag information of each family. Next, observe the distribution information of target. You need to filter out the data of parentecso1 = = 1, parentecso = 1, which means that the family member is the head of a family (a family has only one head of a family). Therefore, the target of the family member with parentecso = 1 is the target of the whole family.

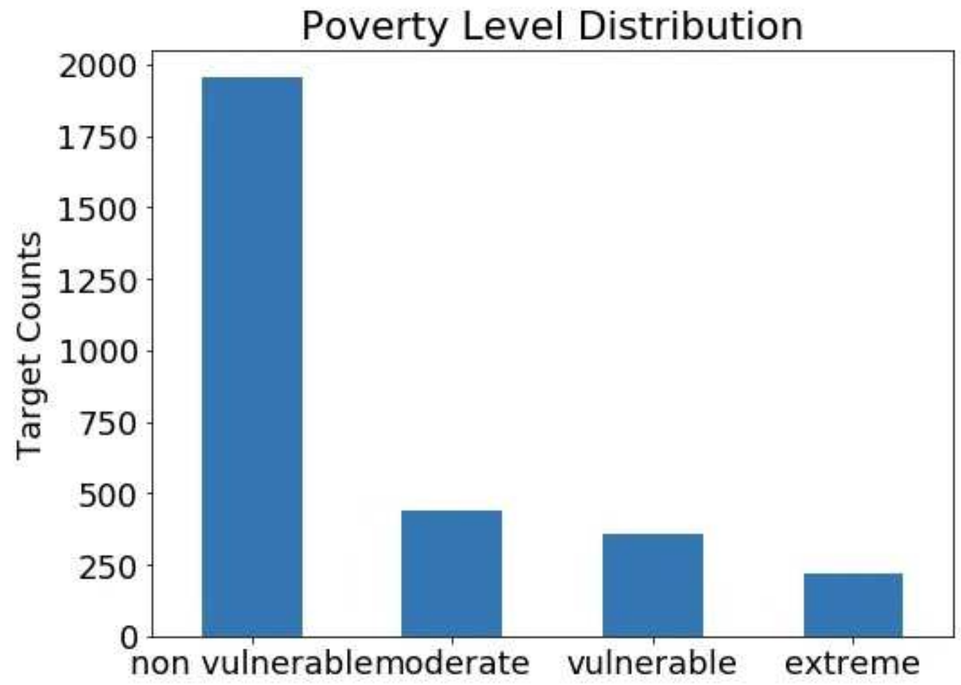


Figure 2: poverty level distribution

Addressing Wrong Target

The normal situation is: if the target of a family is extreme, that is to say, the target of the member who is the head of the family is extreme, then the target of all members in the family should be the same extreme. However, in some abnormal situations, members belonging to the same family have multiple different targets, which may be caused by human or other factors.

First, find families with different family member labels. Then, we can correct all the label differences by assigning the label of the head of household to everyone in the same household. For a family without a head of household, we don't need to worry. If a family does not have a head of household, then there will be no real label and no training will be conducted with any family without a head.

Missing Value Handling

Only look at the existing data, with the most cases as a reference.

Education,Labels,Gender

Next, let's look at the relationship between the education level of different age groups and the poverty situation through the box chart.

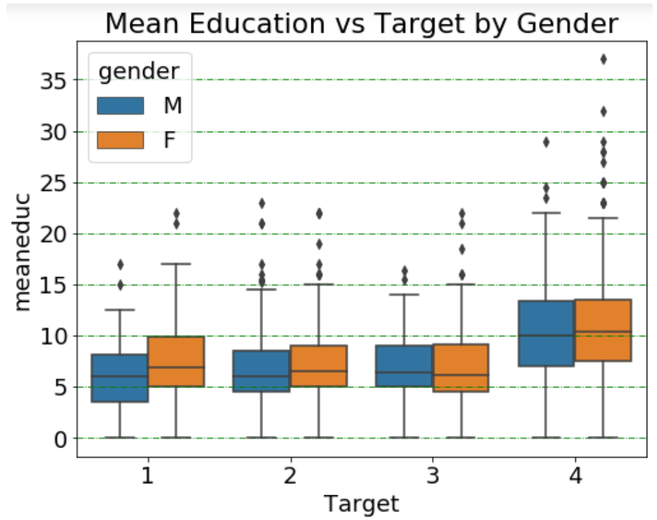


Figure 3: relation diagram

Pearson correlation coefficient

In statistics, Pearson correlation coefficient is used to measure the correlation between two variables X and y, and its value is between - 1 and 1.

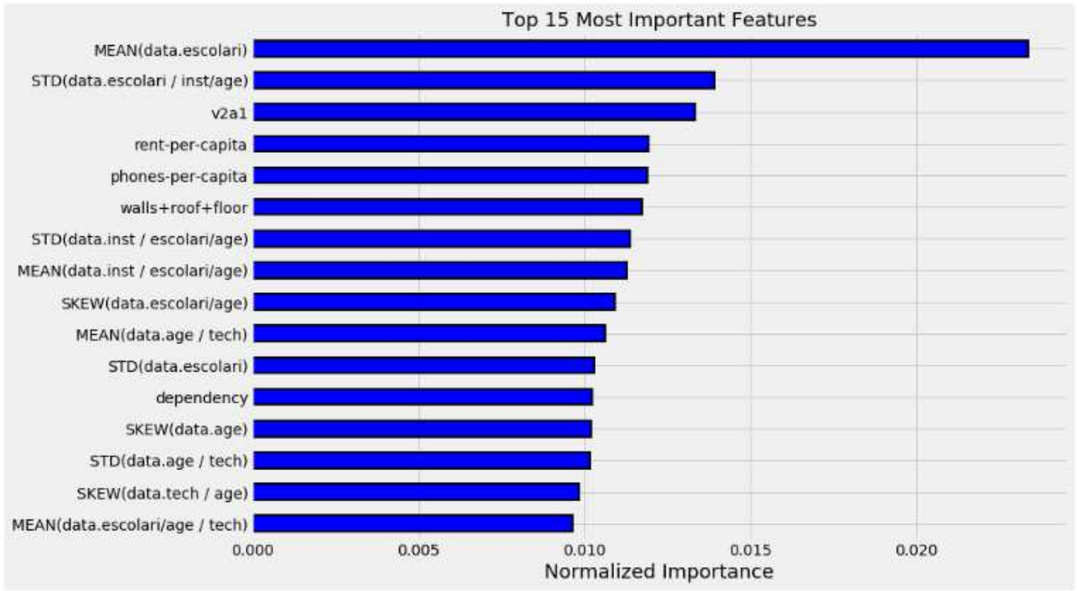


Figure 4: forecast result chart

Method

Method

- RandomForestClassifier

Random forest is a classifier that uses multiple trees to train and predict samples. In machine learning, random forest is a classifier with multiple decision trees, and the output category is determined by the mode of the output category of individual tree.

Result

Forecast Result

From the predicted results, we can see that the number of people belonging to label four, that is, non vulnerable families is the most.

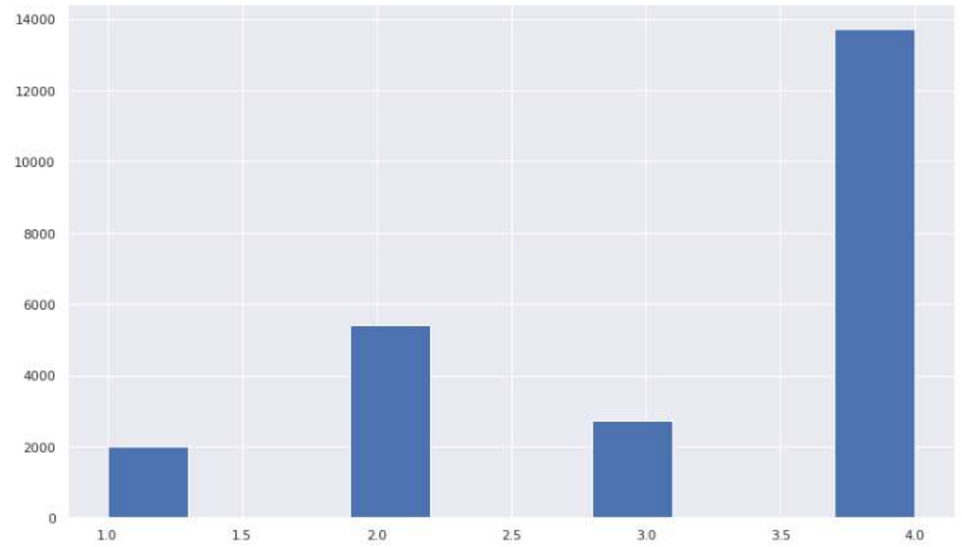


Figure 5: forecast result chart

The Ending