THE REPORT OF FLIP00 FINAL PRESENTATION

ZEBIN JU

Abstract. The report consists of four parts.First, problem description, a detailed description of the problem studied. Second, data analysis. Preprocess the data. Third, a random forest model is used. Fourth, the conclusion, will show the conclusion. The report consists of four parts. (conclusion of problem description data analysis method) first, problem description, a detailed description of the problem studied. Second, data analysis. Preprocess the data. Third, a random forest model is used. Fourth, the conclusion, will show the conclusion.

Contents

Date: 2019-11-29.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

1.1. **Problem Description.** absdefg There are 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. In Latin America, we need to qualify for revenue. It is referred to as agent means test (or PMT). Through PMT, agents use a model, which can analyze the needs of their families by considering the observable family properties, such as the materials of walls and ceilings, or the assets found in their homes. Although this is an improvement, accuracy remains a problem as the population of the region grows and poverty decreases.

2. Problem Analysis

2.1. **Data Information.**

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

Each row represents the data of a family member (a family can be composed of multiple members, only the head of the household is predicted)

Training set: train.csv, 143 columns, ID, target and 141 features

Test set test.csv, including 142 columns, excluding target

The head of household represents a family, and only the head of household is graded

Key fields:

ID: an identification of each sample data 123456789

Idhogar: the unique identification of each family, with the same identification sample belonging to a family

Parentesco1: indicates whether the member is the head of the household. Equal to 1 means the member is the head of the household

Target: poverty of families

1 = extreme poverty

2 = moderate poverty

3 = vulnerable households

4 = non vulnerable households

1 is the most extreme poverty, 2 is the poor, 3 is vulnerable, 4 is not vulnerable

2.2. **Data Analysis.**

2.2.1. *Addressing Wrong Target.*

The normal situation is: if the target of a family is extreme, that is, extreme poverty, that is, the target of the head of household is extreme, then the target of all members of the family should be the same extreme. However, there are some exceptions, the members of 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 has no head of household, then there is no real label.555

2.2.2. *Missing Value Handling.*

Missing value processing I take the following approach. First look at the existing data, with the most cases as a reference. For example, the most common number of tablets in a family is 1, so we fill in the missing places with 1.

2.3. **Data Visualization.**

Data visualization is a scientific and technological research on the visual representation of data. Among them, the visual representation of data is defined as information extracted in some summary form, including various attributes and variables of corresponding information units. Data visualization mainly aims to convey and communicate information clearly and effectively by means of graphical means.

2.3.1. *Data Glance.*

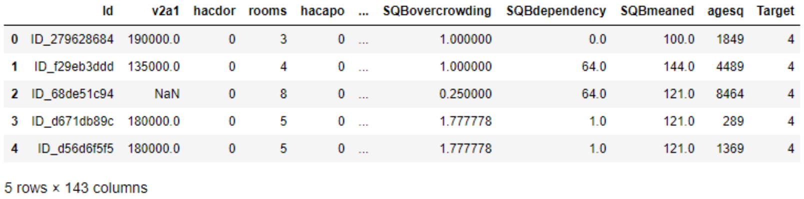
Here, the data list shows the first five rows of data, that is, the situation of five families. The attributes include ID, room number, target tag, etc.

2.3.2. *Histogram.*

Histogram, also known as mass distribution map, is a kind of statistical report map, which is represented by a series of vertical stripes or line segments with different heights Generally, the horizontal axis is used to represent the data type, the vertical axis is used to represent the distribution, and the histogram is the accurate graphic representation of the numerical data distribution.From this histogram, it can be seen that this is obviously a category imbalance, belonging to the fourth category, that is, non vulnerable families have the most people. From this histogram, it can be seen that this is obviously a category imbalance, belonging to the fourth category, that is, non vulnerable families have the most people. 12345899

2.3.3. *Box Diagram.*

Box chart, also known as box chart, box chart or box line chart, is a statistical chart used to display a group of data dispersion information. It is mainly used to reflect the distribution characteristics of the original data, and also to compare the distribution characteristics of multiple groups of data. It can be found from the box chart that the average education level of the extremely poor families with poverty level of 1: extreme is the lowest, both male and female heads of household. With the increase of education level, the poverty level has declined, and it can be found that the average education level of edu and povert is inversely proportional. No matter which level of poor families, the average education level of female heads of household is slightly higher than that of male heads of household.



	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

Sunday is indicated by 6.

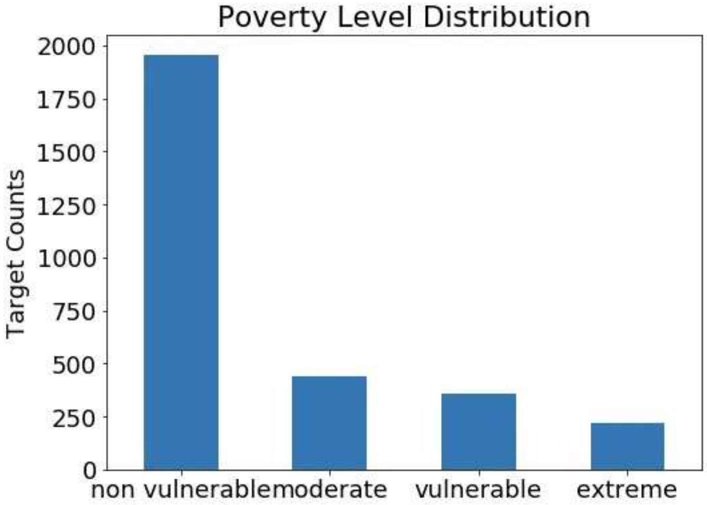


FIGURE 2. Poverty level distribution

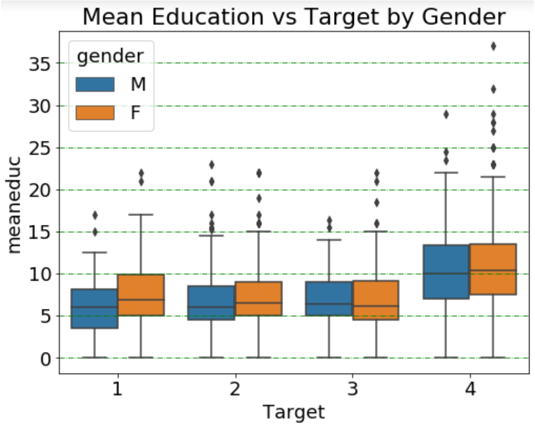


FIGURE 3. Relation diagram

3. METHODS

3.1. **Random Forest.**

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. Random forest is an algorithm that integrates multiple trees through the idea of integrated learning. Its basic unit is decision tree, and its essence belongs to a big branch of machine learning integrated learning method.

4. Resuult

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

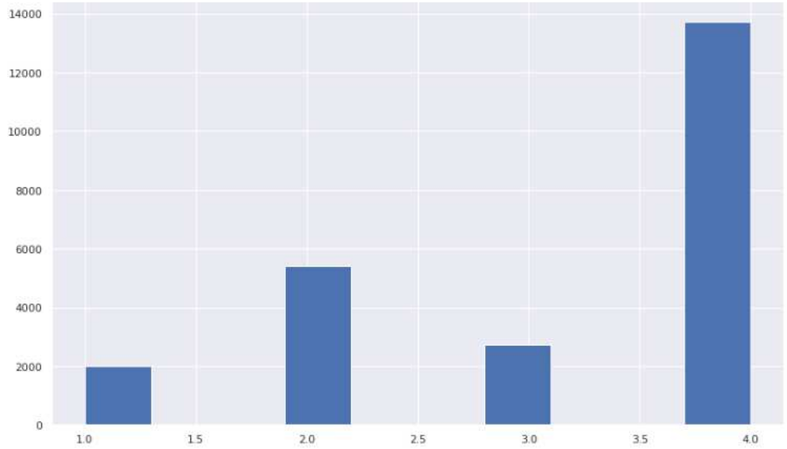


Figure 4. Forecast result chart

5. Conclusion

- The data preprocessing part is very important, which is the basis to solve the later problems. In this part, data processing should be considered comprehensively.
- Random forest is a very flexible and practical method, which has excellent accuracy and can run effectively on large data sets.
- Drawing is also very important.SO you can clearly see the relationship between data.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy