****

# Income Qualification of Costa Rica

Vibhuti Mayekar

Date: 13/07/2020

Contents

[Income Qualification of Costa Rica 1](#_Toc45570290)

[4](#_Toc45570291)

[Description 4](#_Toc45570292)

[Data Exploration: 5](#_Toc45570293)

[Indiviuals: 6](#_Toc45570294)

[Household: 6](#_Toc45570295)

[Overcrowded: 7](#_Toc45570296)

[Average educated people and poverty level: 8](#_Toc45570297)

[**Finding the head of the household:** 8](#_Toc45570298)

[Comparing the no. of people with the Target: 9](#_Toc45570299)

[Finding which Hogar(Place) population: 10](#_Toc45570300)

[Region and Target: 10](#_Toc45570301)

[Area: 10](#_Toc45570302)

[Comparing location with the Target variable: 11](#_Toc45570303)

[Tipovivi(Type of house): 11](#_Toc45570304)

[Comparing the Target with the ‘escolari’ (years of education): 12](#_Toc45570305)

[Data Description 12](#_Toc45570306)

[Integers 13](#_Toc45570307)

[Float 13](#_Toc45570308)

[Categorical 14](#_Toc45570309)

[Count how many null values are existing in columns. 14](#_Toc45570310)

[Dealing with null values: 14](#_Toc45570311)

[Looking at the variables 'v2a1' 15](#_Toc45570312)

[Null value of ‘v18q’ 15](#_Toc45570313)

[Null values in ‘Rez\_escolari’: 15](#_Toc45570314)

[Meaneduc: average years of education for adults (18+): 16](#_Toc45570315)

[Check if there is a house without a family head. 16](#_Toc45570316)

[Set poverty level of the members and the head of the house within a family. 16](#_Toc45570317)

[Check if there are any biases in your dataset. 16](#_Toc45570318)

[Check whether all members of the house have the same poverty level. 17](#_Toc45570319)

[Data Engineering 17](#_Toc45570320)

[Predict the accuracy using random forest classifier. 17](#_Toc45570321)

[Kfold: 18](#_Toc45570322)

[Checking the score using 100 trees 18](#_Toc45570323)

[Important feature: 18](#_Toc45570324)

[Final Interpretation: 18](#_Toc45570325)

# A picture containing grass, outdoor, building, train Description automatically generated

# Description

Identify the level of income qualification needed for the families in Latin America.

Problem Statement Scenario:

Many social programs have a hard time ensuring that the right people are given enough aid. It is tricky when a program focuses on the poorest segment of the population. This segment of the population cannot provide the necessary income and expense records to prove that they qualify.

In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to

classify them and predict their level of need.

While this is an improvement, accuracy remains a problem as the region’s population grows and poverty declines.

The Inter-American Development Bank (IDB)believes that new methods beyond traditional econometrics,

based on a dataset of Costa Rican household characteristics might help improve PMT’s performance.

Following actions should be performed:

* Identify the output variable.
* Understand the type of data.
* Check if there are any biases in your dataset.
* Check whether all members of the house have the same poverty level.
* Check if there is a house without a family head.
* Set poverty level of the members and the head of the house within a family.
* Count how many null values are existing in columns.
* Remove null value rows of the target variable.
* Predict the accuracy using random forest classifier.
* Check the accuracy using random forest with cross validation.

# Data Exploration:

The Target variable is categorized into 1,2,3 and 4, having poverty level as:Extreme, Moderate, Vulnerable, and Non-Vulnerable, respectively.

#### 

Figure

#### From the above figure we can say that the poverty level is the highest i.e. 62% in the level 4 which ‘Non-vulnerable’. This also indicates that data is imbalanced. We are dealing with the imbalanced class problem.

#### The extreme class is very less that is 8%.

## Indiviuals:

#### Distribution of male and females in the house. This indicates, that there is equal no. of males and females in the household.

#### A screenshot of a cell phone Description automatically generated

Figure

#### 

## Household:

#### We will compare the size of the household and no. of persons in the household. The plot says that they both correlate each other. This is very obvious, if the no. of person in the house increase so does the size of the household.

#### A screenshot of a cell phone Description automatically generated

Figure

## Overcrowded:

This figure shows that more no. of people in less bedrooms house, make it overcrowded.

A close up of a device

Description automatically generated

Figure

* The extreme poverty level is only up to the households having 8 rooms in total.
* The household having 9 rooms, has highest vulnerable poverty level, because of overcrowd.
* Household having 10, 11 rooms are very much non vulnerable.

A close up of a pencil

Description automatically generated

Figure

## Average educated people and poverty level:

#### Comparing the ‘Meaneduc’ and ‘Target’ variable, we got to know that, people in the level 4, which is non-vulnerable, are more educated.

#### 

Figure

#### Finding the head of the household: The no. of head in household marked as poverty level 4 is high. In the poverty level 1, the no, there is a smaller number of head in the household.

#### A picture containing screenshot Description automatically generated

Figure

## Comparing the no. of people with the Target:

#### We can see that, at poverty level 3, the no. of people staying in the household are more.

#### 

Figure

## Finding which Hogar(Place) population:

#### We can see that the population is more in the Central region.

A screenshot of a cell phone

Description automatically generated

Figure

## Region and Target:

‘Central’ and ‘Chortega’ are more under non- vulnerable region.

A picture containing drawing

Description automatically generated

Figure

## Area:

The urban area seems to be more under the 4th poverty level which non-vulnerable. The rural area is on 3rd level, which is ‘Vulnerable’.

**A screenshot of a cell phone

Description automatically generated**

Figure

## Comparing location with the Target variable:

We can see that; non-vulnerable poverty level people stay in Central region in highest number. Also, Central is the region which has mixed crowed, i.e of all poverty level higher. That I obvious, because central has the most population.

A screenshot of a cell phone

Description automatically generated

Figure

## Tipovivi(Type of house):

Looking at the figure, we can say that house condition in ‘Precarious’ and ‘assigned’, their poverty level is moderate. While house which fully owned, or instalments, or rent, their poverty level is non-vulnerable.

A screenshot of a cell phone

Description automatically generated

Figure

## Comparing the Target with the ‘escolari’ (years of education):

People staying in nonvulnerable poverty lane, are more educated.

A screenshot of a cell phone

Description automatically generated

Figure

# Data Description

There are 143 entries, out of which integers=130, float = 64, and categorical = 5.

## Integers

A close up of a logo

Description automatically generated

Figure

## Float

A screenshot of a cell phone

Description automatically generated

Figure

## Categorical

A picture containing drawing

Description automatically generated

Figure

# Count how many null values are existing in columns.

## Dealing with null values:

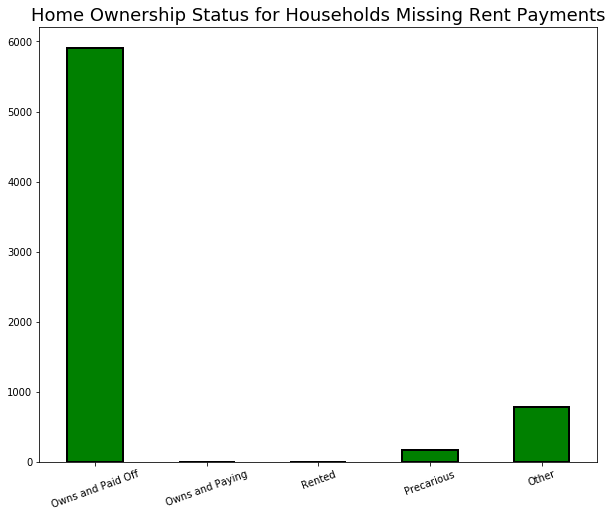
There are null values in the variables: ‘v2a1’(Monthly rent payment)=6860, ‘v18q1’(no. of tablets owned)=7342, ‘rez\_esc’ (Years behind school)=7928, ‘meaneduc’ (average years in school)=5, ‘SQBmeaned’ (square of the mean years of education of adults (>=18) in the household)=5.

A close up of a logo

Description automatically generated

Figure

Looking at the variables 'v2a1' and 'tipovivo1', 'tipovivo2', 'tipovivo3', 'tipovivo4', 'tipovivo5' which is "Own and fully paid", "own, paying in installments", "rented", "precarious", “other(borrowed/assigned)” respectively. ‘v2a1’ has null values in all those places, where the household is owned. So we will replace the null value with 0.



Figure

Null value of ‘v18q’, which is no. of households having tablets. Looking at the variable ‘v18q1’, which is total no. of tablets in the households, it is observed that, ‘v18q’ has null values at those places where ‘v18q1’ is 0. So, we can replace null values of ‘v18q’ with 0.

Null values in ‘Rez\_escolari’: We will compare the rez\_esc with the age of the members in the household.

“train.loc[(train['rez\_esc'].isnull() & ((train['age']>7) & (train['age']<17)))]['age'].describe()”

count 1.0

mean 10.0

std NaN

min 10.0

25% 10.0

50% 10.0

75% 10.0

max 10.0

Name: age, dtype: float64

It is observed that, wherever the age is 10, there is a null value. So we will replace that value with 0.

## Meaneduc: average years of education for adults (18+):

lets compare it with the no level education ‘instlevel0’. These columns are correlated.

train.loc[train['meaneduc'].isnull()]['instlevel1']

1291 0

1840 0

1841 0

2049 0

2050 0

Name: instlevel1, dtype: int64

Similarly, we will replace ‘meaneduc’ and ‘SQBmean’ missing values with 0.

# Check if there is a house without a family head.

There are 15 households having no heads.

# Set poverty level of the members and the head of the house within a family.

There are 0 households where the family members do not all have the same target.

# Check if there are any biases in your dataset.

Extreme is the smallest. So data is biased.

A screenshot of a cell phone

Description automatically generated

Figure

# Check whether all members of the house have the same poverty level.

There are 85 households where the family members do not have the same target or poverty level.

# Data Engineering

Finding all the variables which are redundant to each other and if necessary, remove them.

1. Removed all 'SQBescolari', 'SQBage', 'SQBhogar\_total', 'SQBedjefe', 'SQBhogar\_nin', 'SQBovercrowding', 'SQBdependency', 'SQBmeaned', 'agesq'. They are average of existing variables. They can be easily removed.
2. Checking the redundant variables: We can remove variables like 'tamhog', 'hogar\_total', 'r4t3', since they are highly corelated to each other.

A screenshot of a cell phone

Description automatically generated

Figure

1. Dropping one of the columns of ‘male’ and ‘female’. Keeping one will suffice. We will drop ‘male’.
2. Similarly, between ‘area1’ and area2’, we will drop area2.
3. We will also delete 'Id', 'idhogar'.
4. Final shape of data: (9557, 127)

# Predict the accuracy using random forest classifier.

After performing the random classification on the train set data, we got following accuracy.

print(accuracy\_score(y\_test,y\_predict))

print(confusion\_matrix(y\_test,y\_predict))

print(classification\_report(y\_test,y\_predict))

0.9252092050209205

[[ 130 2 2 23]

[ 3 274 4 36]

[ 5 7 179 42]

[ 0 12 7 1186]]

precision recall f1-score support

1 0.94 0.83 0.88 157

2 0.93 0.86 0.90 317

3 0.93 0.77 0.84 233

4 0.92 0.98 0.95 1205

accuracy 0.93 1912

macro avg 0.93 0.86 0.89 1912

weighted avg 0.93 0.93 0.92 1912

* Accuracy that we got here is 93%

Kfold: Lets check with the kfold cross validation.

We randomly use k= 5 and got following accuracy:

[0.93514644 0.93043933 0.92569335 0.9314495 0.9314495 ]

93.08356268159017

It is same which we got thru Random classifier.

## Checking the score using 100 trees

[0.94246862 0.94979079 0.94557823 0.94243851 0.94976452]

94.60081361157272

Almost same.

## Important feature:

Following figure indicates, that ‘Average years of schooling’, ‘dependency’, ‘overcrowding are the important features and the rest follow.

A close up of a logo

Description automatically generated

Figure

# Final Interpretation:

* [Figure 1](#_Data_Exploration:) shows that the highest population is under ‘non-vulnerable’ poverty level. Which shows that the data is biased. Also, we need more data to come to conclusion. Nevertheless, after performing analysis on the given data, we can say that the bank shall focus only on those ‘extreme’ poverty level population. They are the much needy people.
* [Figure 10](#_Region_and_Target:) shows us that the region which is more vulnerable and needy are ‘PacAfico Central’, ‘Brunca’, ‘Huetar AtlÃ¡ntica’, ‘Huetar Norte’. This does not mean that central and Chortega does not have ‘extreme’, ‘moderate’, and ‘vulnerable’ poverty levels.
* Looking at the figure, we can say that house condition in ‘Precarious’ and ‘assigned’, their poverty level is moderate. While house which fully owned, or instalments, or rent, their poverty level is non-vulnerable. Figure
* [Figure 5](#_Overcrowded:) shows that the ‘Vulnerable’ poverty level is in the house having 9 bedrooms. The reason behind this is overcrowding. Probably the population to bring under control will be the solution for this.
* ‘Meaneduc’ average years of schooling plays a vital role in standard of living of the people.
* People with poverty level as ‘Non vulnerable’ has highest education. [Figure 6](#_Comparing_the_Target)
* After using Random Forest classifier on 127 variables, we got 93% accuracy. So, our model can work well on new data with these variables.
* However, data is biased, the only solution to get the model learn more, is to increase the data. The given 9557 rows are not sufficient. To increase the model efficiency, there is a need to increase the amount of data with rows.
* However the Inter-American Development Bank, shall consider the parameters like ‘Average education’, ‘Dependency’, ‘Overcrowding’ in to consideration for future income qualification study.