Income Qualification Project Writeup

Here, I follow the steps outlined in the project description. I spend more of my time examining rows 1086-1100 because they show many different values of the target variable close together.

Identify the output variable

The output variable must be "Target". It only appears in the training dataset, not the testing one. Target is a number between 1 and 4, which classifies the household. The value is lower for households with less need and higher for household with greater need.

Understand the type of data

In the Target column, we use numbers from 1 to 4 to categorize each household. The classification depends on the values in the other columns. Almost every column in the dataset contains numerical data. Only the columns Id, v18q1, rez_esc, idhogar, dependency, edjefe, and edjefa contained categorical data or a mix of numerical and categorical data.

In the case of edjefe and edjefa, yes = 1 and no = 0. Upon inspection, the same applies to dependency. If dependency = no, then dependency = 0 and there are no dependents. If dependency = yes, then dependency = 1 and the dependency ratio is 1.

This dataset consists of 9557 individuals. The individuals are kept track of with the Id column. The households that the individual belongs to is kept track of with the idhogar column.

Columns by group:

Based on the list of columns given in the PDF file, I can say the following about the data:

Columns pertaining to home ownership status

Column 2: Monthly rent payment. Useful.
Columns 116-120: Household ownership status. Useful.

Columns pertaining to number of rooms and overcrowding

Columns 3-6: Deals with number of rooms and overcrowding. Useful, but redundant.
Column 114: Number of bedrooms.
Column 115: Overcrowding. # persons per room.

Columns pertaining to number and type of people in home

Columns 10-20, 23: Number of people in home by age group. Much of this is redundant, but it is useful.
Columns 84-95: What sort of people live in this house? These columns are most useful for determining whether there is a family head, whether the family head has a spouse, and if there are any people living in the home that are not part of the immediate family. The presence of grandparents, grandchildren, in-laws, and non-family members indicates greater need because most households don't have all these people.

Column 74: It looks like a boolean that indicates if disabled people are present.
Column 75: It looks like a boolean that indicates if males are present.
Column 76: It looks like a boolean that indicates if females are present.

Columns 97- 100: Number of people in age group. In truth, this set of characteristics is more useful than most of the other people-related columns because we are dealing strictly with finances. This can replace most of the people-related columns for analysis.
Column 101: Dependency. This is a ratio based on people in age groups. This might be more useful than columns 97-100 and by extension the rest of the people-related columns.

Column 133: Age. But age of what? We already have the age of the people living here.

Columns pertaining to house material quality

Columns 24-31: Primary material on outside wall. Slightly useful. Can mostly be ignored, except for the case of asbestos.
Columns 32-35, 37: Floor material type. Not very useful.
Column 36: Is floor present? Useful.
Columns 38-41: Roof material type. Not very useful.
Column 42: Is roof present? Useful.
Columns 65-67: Wall quality. Useful.
Columns 68-70: Roof quality. Useful.
Columns 71-73: Floor quality. Useful.

Columns pertaining to necessary facilities and appliances for good life quality

Column 7: Owns a refrigerator. Useful.
Column 43-45: Water provision type and check if water provisions are present. Very useful to know whether water is present.
Column 46-49: Where electricity comes from. Very useful.
Columns 50-54: Type of toilet. Is toilet present? Useful.
Columns 55-58: Where does the kitchen get power from? Useful.
Columns 59-64: How is rubbish disposed of? Useful.

Columns pertaining to education

Columns 21-22: Years of schooling. Redundant, but useful.
Columns 105-113: Duration of education. Mostly redundant because columns 102-104 cover the education of the heads of household.

Columns pertaining to household heads

Columns 102-104: Education of family heads and adults. Useful because these people are the breadwinners.
Columns 77-83: Marriage status. This is can affect the household dynamics and the total income for the family.

Columns pertaining to home location

Columns 125-130: Region of country household is in. Not useful except if I knew which regions tended to be richer and poorer.
Columns 131-132: Urban or rural home? Rural homes tend to be poorer and less infrastructure. Useful.

Columns pertaining to electronics

Columns 121-124: Electronics owned by household. Not a strong indication of need. Not useful except to indicate whether a household has the finances to buy such things.
Columns 8-9: Owns a tablet. Redundant and not useful.

Columns whose use I can't determine

Columns 134-142: These are just squares of previous values. We don't need these.

Check if there are any biases in your dataset

It is difficult to determine biases in the dataset because what appears like a bias may actually be easily explained by examining all of the other columns. Characteristics about individuals that are not objective measures of their need, such as gender, age cutoff, and home construction material.

Those who own their homes (tipovivi1 = 1) have lower target values and are thus given lower need.

When overcrowding = 2, then the target values are higher, between 2 and 4. If overcrowding > 2, then target is usually 4. If overcrowding < 1, then the target values are low, usually 1. We can say that less overcrowding indicates less need.

Interestingly, the household size, hhsize, does not strongly correlate with the value of the target. It's slight, but it appears that the more females a household has, the lower the target value. This fact is true for homes with similar hhsize values. Apparently women are given lower priority on the scale for need. Looking at the values of r4h3, r4m3, and parentesco1 with similar hhsize, the value of the target can vary with number of people of each gender.

However, age can also play a factor. Examining rows 1087 and 1088, it appears that men are given more priority, but examining hogar_mayor shows that the man in question is over 65 years old, giving him greater need and making his target value higher. However, the woman is 60 years old and the man is 77, so they are both seniors, but being over the age of 65 makes a difference. **That particular 65 number is a cutoff**.

Looking are rows 1087 and 1088, which are homes ee3d80cb6 and a0695cb68 respectively, we see that 1087 has a socket exterior on his home, which could include harmful asbestos. Further, there may be a bias that indicates that wooden exteriors and interiors are luxurious. The woman living in row 1088's a0695cb68 home has a wooden home exterior and wooden flooring and has lower need than row 1087's man. **This might be a result of the possible abestos, but it might be a bias in favor of wooden structuring.**

Looking at rows 1091 though 1093, homes of only adults between 19 and 65 years old are given higher priority status, even if they are not disabled and dependency is 0. Examine hogar_adul.

However, I noticed that cielorazo = 0 for these rows, indicating that the people living in 8e284abd5 have no roof. This is a huge problem and not a bias.

In short, I can say that the age cutoff of 65 increases one's need, even if the individual is close to 65, but under that age. The judgment of home material might be indicative of a bias, but it might also be concern over asbestos in the home. Asbestos can cause cancer, so increasing need based on its presence is not illogical. It is not clear is any bias exists since logical explanations exist for each potential bias.

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

Using a for loop, I determined that the following 85 homes have family members that DO NOT have the same poverty level:

'4b6077882', '6833ac5dc', '43b9c83e5', '5c3f7725d', '0f9494d3a', 'daafc1281', '73d85d05d', 'bcaa2e2f5', '44f219a16', 'efd3aec61', '3c6973219', '0511912b6', 'f006348ed', 'a20ff33ba', '5e9329fc6', 'e65d4b943', '42ec8bef5', '6bcf799cf', '26b3a0f41', '4dc11e11f', '594d3eb27', 'd9b1558b5', '7ea6aca15', '8bb6da3c1', '3df651058', '811a35744', '2cb443214', 'bcab69521', '694a0cbf4', '3fe29a56b', '636330516', '288579c97', '15a891635', '6a389f3de', 'a3288e6fa', '4e19bd549', '80a66379b', '5c6f32bbc', '932287f5d',

'bd82509d1', '614b48fb7', '46af47063', '6c543442a', '410194c8b', '417865404', 'f7b421c2c', '67ad49822', '17fb04a62', 'c38913488', '513adb616', 'dfb966eec', '30a70901d', '18832b840', '7c57f8237', 'c13325faf', '54118d5d9', '0f3e65c83', '03f4e5f4d', '8ae3e74ca', '309fb7246', '09e25d616', '564eab113', '8242a51ec', '0172ab1d9', 'a94a45642', 'be91da044', '50e064ee8', '4c2dba109', '7ad269eef', '3c73c107f', '55a662731', 'e17b252ed', '078a0b6e2', '28893b5e7', 'd64524b6b', '2c9872b82', 'f94589d38', '8420bcfca', '71cd52a80', '654ef7612', 'cc971b690', '7e9d58c5c', 'e235a4eec', 'c7ce4e30c', '9bbf7c6ca'

Check if there is a house without a family head

With another for loop, I determined that there are 15 families without family heads. They are:

'09b195e7a', '896fe6d3e', '61c10e099', '374ca5a19', 'bfd5067c2', '1367ab31d', '6b1b2405f', 'f2bfa75c4', '03c6bdf85', 'ad687ad89', 'b1f4d89d7', 'c0c8a5013', 'a0812ef17', 'd363d9183', '1bc617b23'

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

I had no idea how to do this. Should I have found the mode of the target value for each household and assigned it to each family?

Count how many null values are existing in columns

There are 22140 null values in the training dataset.

Remove null value rows of the target variable

I used the following lines to accomplish this goal:

trainDF["Target"].isnull().any()

and

trainDF["Target"].isna().any().

With the above code, I proved that there are no null rows in the target variable. Nothing needs to be removed.

Predict the accuracy using random forest classifier

The code for this can be viewed in the notebook. My approach was to split the training dataset into an input dataset, X, and an output dataset, y.

To make X, I took the training dataframe and removed the Id and idhogar columns because they are categorical data that can't be used here. I removed the target column because that is not input data. It is output data. I replaced all the instances of "no" with 0 and "yes" with 1. I replaced all the NaN values in the training dataset with 0, the logic for which is explained below

I set y to be the target column, "Target".

NaN values

The columns with NaN values are:

v2a1, v18q1, rez_esc, meaneduc, SQBmeaned

The following can be stated about each column:

v2a1 is monthly rent payment and is NaN when rent is not needed. We can put zero here.
V18q1 is NaN when no tablets are owned, so we put 0.
rez_esc is NaN when the individual is not behind in school. Put 0.
meaneduc is NaN when the individual has no education. Put 0.
SQBmeaned is square of meaneduc, so put 0.

Thus, all NaN values in the training dataset will become 0.

Results

After running the random forest classifier code, the accuracy score comes out to be 0.8971408647140865, which is about 0.8971 or 89.71%. This is a good score because it is close to 90% accurate.

Check the accuracy using random forest with cross validation

I used the same inputs X and y from the previous random forest classifier code. With cross-validation, I used 10 splits.

The accuracy of the random forest classifier is 60.916569913908305%, which is approximately 60.92%. This is not a good score because it comes nowhere near 100% accurate. In fact, it is closer to 50% than 100%.

Conclusion

Since the cross-validation random forest model came out to be only 60.91% accurate, I can conclude that the model needs improvement. Many of the columns are redundant. There are too many columns containing values about how many people live in home, how many people are educated, what electronics are owned, the gender of the individuals, the age of the individuals, and others. Removing these columns and condensing the data down to the most useful metrics without redundancies may improve the accuracy of the model.