Predicting Poverty

James Spalding, Ben Bronoski, Matt Nowell, Yaya Barrow

2024-12-07

Introduction

There is a constant struggle to ensure people are being given access to the correct amount of aid they need to survive. Some programs target some of the poorest populations to ensure they are being properly taken care of. Unfortunately, some of these poorer communities are unable to correctly and accurately document that they qualify for the amount of aid they tend to need. The goal of this project is to take observable attributes of a given household and bucket them into different poverty levels.

The data provided includes 142 predictor variables, with a decent spread between categorical, numeric, and binary values. Our response, **Target**, is a categorical variable with 4 levels, and each row represents an observed individual. Below is a list of variables with some omitted or modified for clarity:

Variable	Type	Description
v2a1	Numeric	Monthly rent payment
hacdor	Numeric	Overcrowding by bedrooms
rooms	Numeric	Number of rooms in house
hacapo	Numeric	Overcrowding by all rooms
v14a	Binary	Has bathroom in household
refrig	Binary	Has refrigerator in household
v18q1	Numeric	Number of tablets household owns
r4t1	Numeric	Persons younger than 12 years
r4t2	Numeric	Persons older than 12 years
escolari	Numeric	Years of schooling
rez_esc	Numeric	Years behind in school
hhsize	Numeric	Household size
pared	Categorical	Wall material
piso	Categorical	Floor material
techo	Categorical	Roof material
cielorazo	Binary	Presence of ceiling in home
abastagua	Categorical	Home water source
elec	Categorical	Home electricity source
sanitario	Categorical	Home plumbing type
energcocinar	Categorical	Home kitchen type
elimbasu	Categorical	Home waste disposal type
epared	Numeric	Wall quality
etecho	Numeric	Roof quality
eviv	Numeric	Floor quality
dis	Binary	Individual is disabled
gender	Binary	Individual gender
estadocivil	Categorical	Individual civil status
parentesco	Categorical	Relation to head of household
hogar_nin	Numeric	Individuals under 19

hogar_adul	Numeric	Individuals between 19 and 65
hogar_mayor	Numeric	Individuals >65
dependency	Numeric	Ratio of dependents/independents
edjife	Numeric	Education of head of household (male)
edjifa	Numeric	Education of head of household (female)
meaneduc	Numeric	Mean years of education in household
instlevel	Categorical	Highest form of education achieved
bedrooms	Numeric	Number of bedrooms
tipovivi	Categorical	House status (rent, own, etc)
computer	Binary	Presence of household computer
television	Binary	Presence of household TV
lugar	Categorical	Region
area	Binary	Urban/Rural
age	Numeric	Individual age
Target	Categorical	Household poverty level

Furthermore, our response variable **Target** has the following categories which we will attempt to classify households into:

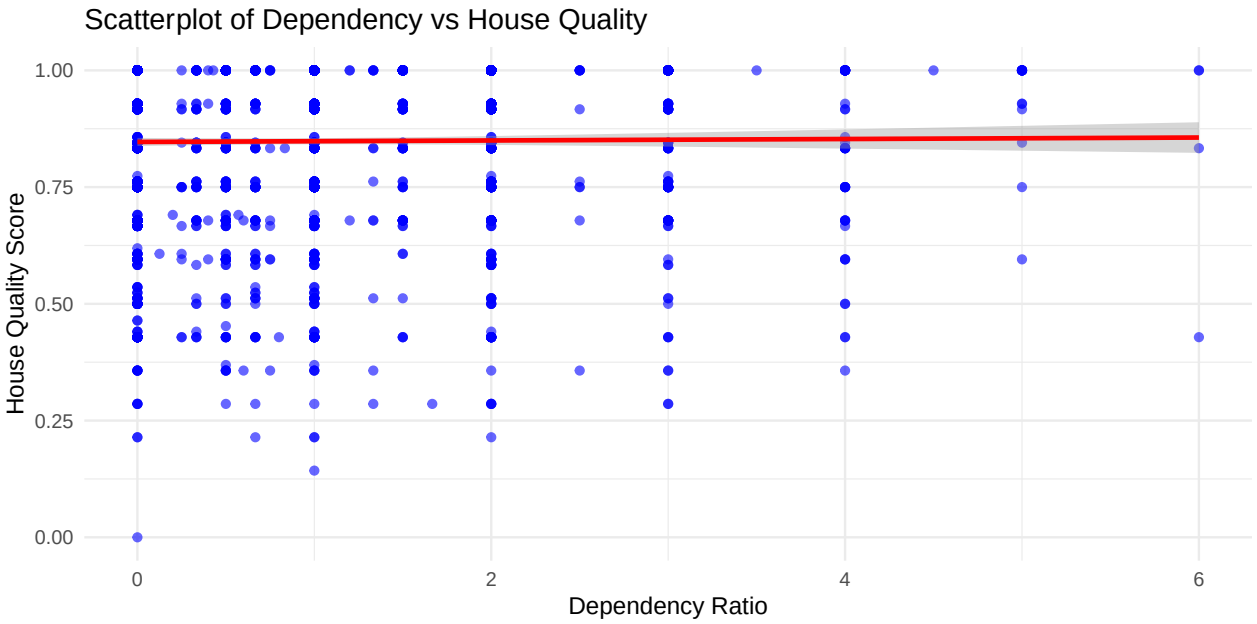
Level	Description	Count
1	Extreme poverty	211
2	Moderate poverty	420
3	Vulnerable	339
4	Non-vulnerable	1849

As shown by the counts, this is a very imbalanced dataset and measures will need to be taken to account for this.

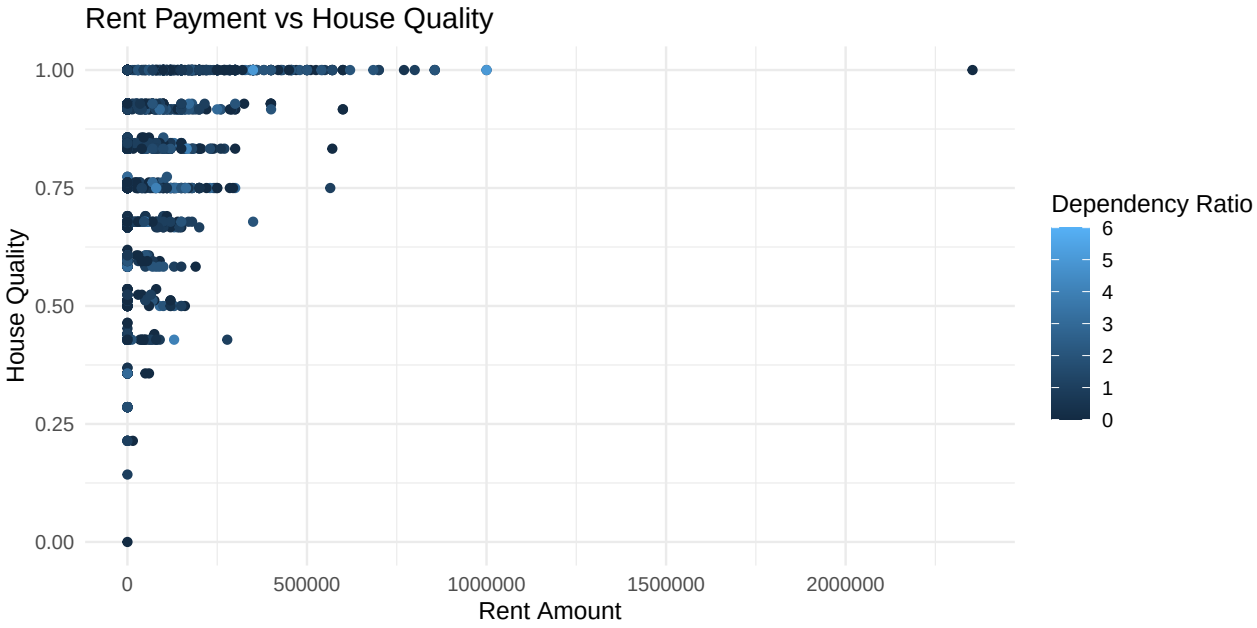
Data Cleaning and Exploration

There are several variables, such as SQBescolari, SQBage, SQBhogar_total, SQBedjefe, SQBhogar_nin, SQBovercrowding, SQBdependency, SQBmeaned, agesq, that are either irrelevant in the analysis, or hold little value to the outcome of the analysis. These variables are squares of other existing variables so they are removed from the dataset to avoid any colinearity. Additional data cleaning steps address missing values, standardize data formats, and compute new variables to ensure the dataset is ready for analysis. Missing values are filled with logical defaults or calculated averages. One example of this is replacing v2a1 (rent payment) and v18q1 (tablets owned) with zero, or using the mean education level for specific age groups to impute meaneduc. Categorical variables, like edjefa and edjefe, are converted into binary numeric formats. Dependency ratios are recalculated at the household level using age group distributions, and a binary indicator for school attendance is created.

Once these steps have been executed, new variables are made to address the ambiguity of some of the existing features. Some of these variables create household-level summaries, such as counts of individuals in school, children behind in school, and disabled members, and identifies households with non-family members. A filter is applied to the dataset to include only heads of households and removes redundant or constant features. Housing quality is quantified through composite scores for interior, exterior, and overall quality based on materials and utilities. To visualize household dependencies, a simple scatterplot was created.



As we can see from this plot, as dependency ratio increases, the house quality seems to have a slightly larger spread which could indicate that having more dependents could have an impact on the quality of a household. Another argument that could be made is that people that have more dependents appear to be in a position to provide for these dependents since there seems to be an upward trend of the house quality score as the dependency ratio increases. Additionally, there could be a relationship between the rent paid and the house quality. To view this, another scatterplot was made.



As we can see in this plot, there seems to be some sort of relationship when rent goes up, the house quality rises with it. There also seems to be a lower dependency ratio as both rent and house quality increase.

Base Model

Our first step is to get a baseline model. Since our response is categorical with 4 levels, we will be using a multinomial logistic regression model. We separated our data with a 70% train/test split and created a model using all variables created and retained in the data cleaning process. The performance of this model is shown below:

Base Model	1	2	3	4
Sensitivity	0.22	0.31	0.05	0.91
Specificity	0.96	0.90	0.97	0.46

As shown, the initial model performs quite well in class 4, decent in classes 1 and 2, and terribly in class 3.

To improve our model, we first attempted elastic net with crossfold validation (10 folds). We initially attempted ridge regression, but noticed that the model was only predicting cases from levels 2 and 4; the two largest groups within the data. We next tried LASSO regression, and had nearly identical results. Finally, in order to rule out elastic net as a viable strategy for our data, we tested every α value between 0 and 1 in intervals of 0.05. Out of these 20 tests, $\alpha = 0.3$ was able to make predictions on levels 1, 2, and 4, $\alpha = 0.7$ was able to make predictions on all 4 classes, and all others were only able to predict on 2 and 4. However, neither $\alpha = 0.3$ nor $\alpha = 0.7$ were able to make accurate predictions, so none of the elastic net models will be used.

We hypothesized that, since the predictors appeared to not be a limiting factor, the class imbalance must be the reason for poor predictions. To account for this, we tested a few different weights on each of the class probability predictions. The weights, along with the performance of their corresponding models, are shown below:

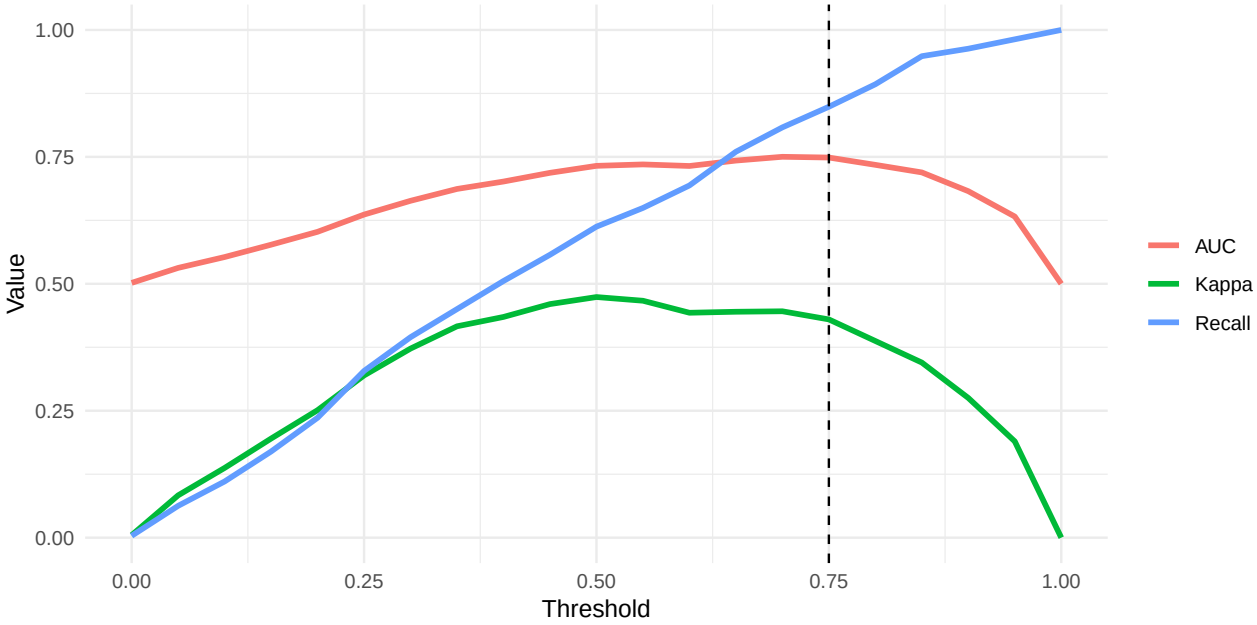
Weight	κ	AUC
None	0.26	0.65
$\frac{1}{\text{prop}}$	0.27	.070
$\frac{1}{\sqrt{\text{prop}}}$	0.31	0.68

As shown, both of these models perform above the unweighted model in both κ and ROC-AUC scores. The trade off, however, is that neither of these models are nearly as accurate at predicting class 4 as the unweighted version.

Multi-Model Approach

Since our weighted model did so much better in predicting classes 1-3, we will continue to use it. However, to improve on its lacking capabilities in predicting class 4, we will introduce a second model which *exclusively* predicts class 4.

For this model, the Target variable has been transformed to a binary classifier as to whether the case falls into class 4 or not. In this model, we want as high of an specificity as we can reasonably get, since false positives won't get a second chance at being properly classified by model 2. Since recall (the amount of true positives divided by the total predicted positives) appears to scale linearly with threshold, we took the highest threshold with an acceptable κ and AUC score. This value ended up being 0.75. The results are shown in the plot below:



Now that we have a threshold selected for the binary model, all we need to do is pass the values that it predicts are not 4 to the original model. The results are shown in the following table:

Final Model	1	2	3	4
Sensitivity	0.50	0.30	0.30	0.68
Specificity	0.88	0.87	0.84	0.80

While the sensitivity for observations in class 4 may have fallen by a decent margin, its specificity rose to compensate. While metrics for class 2 predictions are more or less the same, metrics for class 1 and 3 predictions have seen substantial improvements.

Conclusion