Classifying Household Poverty Levels

Julia Haas, Eva Peters, Sophie Pope, and Maddy Rilling

Introduction

This project aims to create a model to classify a household into one of four poverty levels. To do this we will be looking at information for the head of each household. The main motivation behind this comes from the fact that many poverty stricken households lack the resources or capability to report their income, making it difficult for social programs to properly give out aid. Our model will be using observable household attributes rather than income to classify households into poverty levels. To do this we started by analyzing relationships between predictor variables as well as between predictors and poverty levels. We then came up with our baseline model, and put it through backward elimination to come up with our final model.

Data Exploration

Description of the Data

This data set contains many variables that describe the demographics of both individuals and households. Each row represents an individual, totaling to just over 9,000 rows. Each row includes basic individual demographic information such as age, gender, marital status, and years of education, as well as various household attributes such as the material the house is made of, total bedrooms, if the home is overcrowded, and ownership status. In total, the data started with 142 predictor variables and 1 response variable classifying a person into one of four poverty levels: “extreme poverty,” “moderate poverty,” “vulnerable households,” and “non-vulnerable households.” Looking at the distribution of this variable, about 62.8% of the individuals were classified as non-vulnerable, 12.7% as vulnerable, 16.6% in moderate poverty, and 7.9% in extreme poverty.

Cleaning the Data

To start, we renamed the variables in order for them to make more sense when working with them for data exploration. We then created many categorical variables by carefully grouping binary variables together into a single column, such as grouping different binary floor material variables into a single floor material variable. After that, we filtered our data to include only rows for heads of the households, since we will be making our model based on those types of individuals. We dealt with NA values by examining why they were occurring, and found the main one to be under the number of tablets row if the family did not own a tablet, so we dealt with those missing values by changing them to 0’s. After that 25 NA values remained, so it was reasonable to drop these rows as it is a very small proportion of the data. Finally, we dropped unnecessary columns in our data, which included columns that were now combined into one categorical variable, as well as variables that were about the same except for a few cases, such as number of people in a household versus number of people living in the household, especially since it would be difficult to differentiate between them.

Visualizing Relationships

Figure 1: Ownership of Assets by Poverty Level

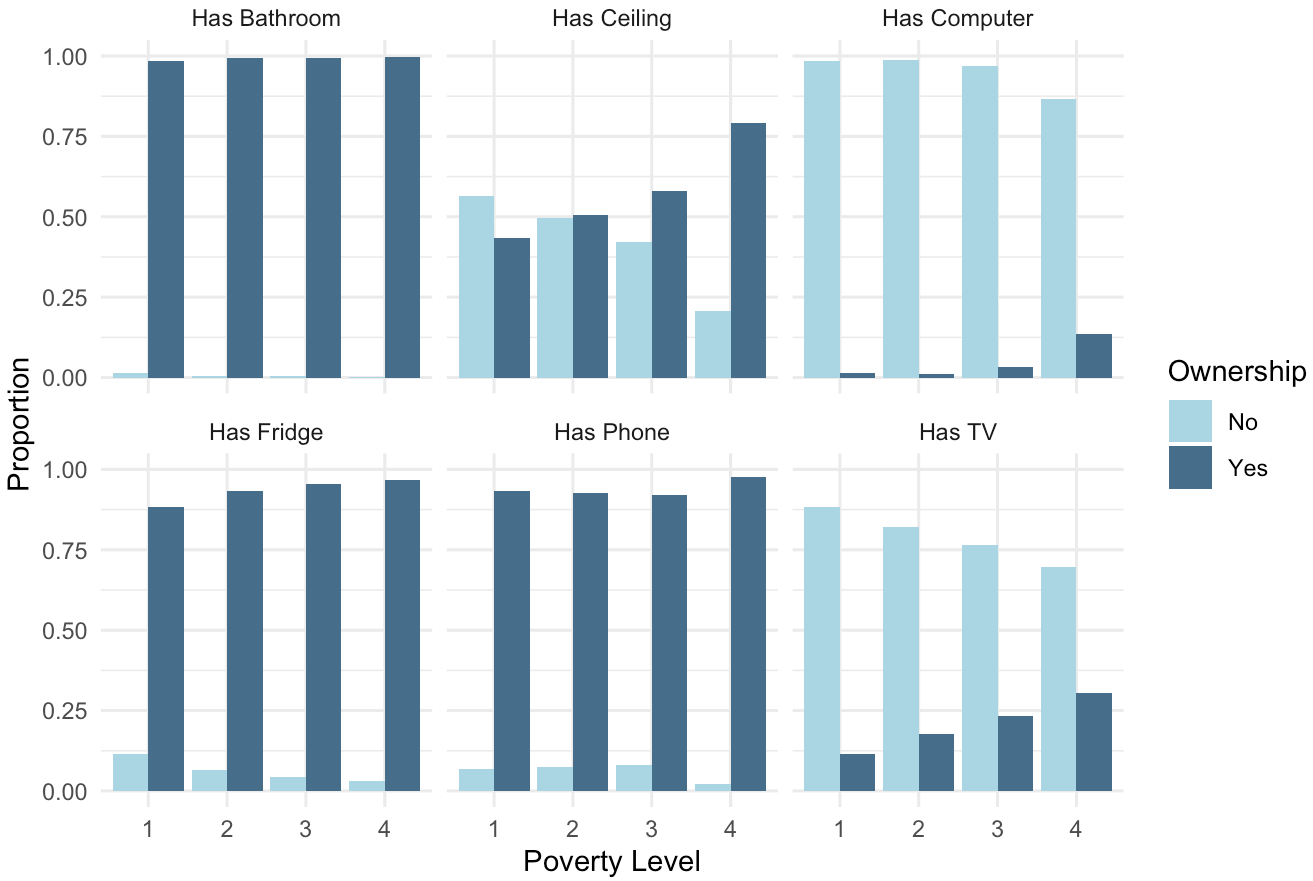

Figure 1 visualizes the relationships between a household's poverty level (1 being extreme poverty and 4 being non-vulnerable households) and six different possible binary predictors. We can see that the more a household is in poverty, the less likely they are to have a TV, computer, ceiling, phone, and fridge. This trend is most apparent for having a TV and having a ceiling. The bar plot visualizing if the household has a bathroom does not have any significant differences for the four levels. This visualization indicates that there are some notable relationships between poverty status and most of the binary predictors displayed above.

Figure 2: Overcrowded Status by Number of Children

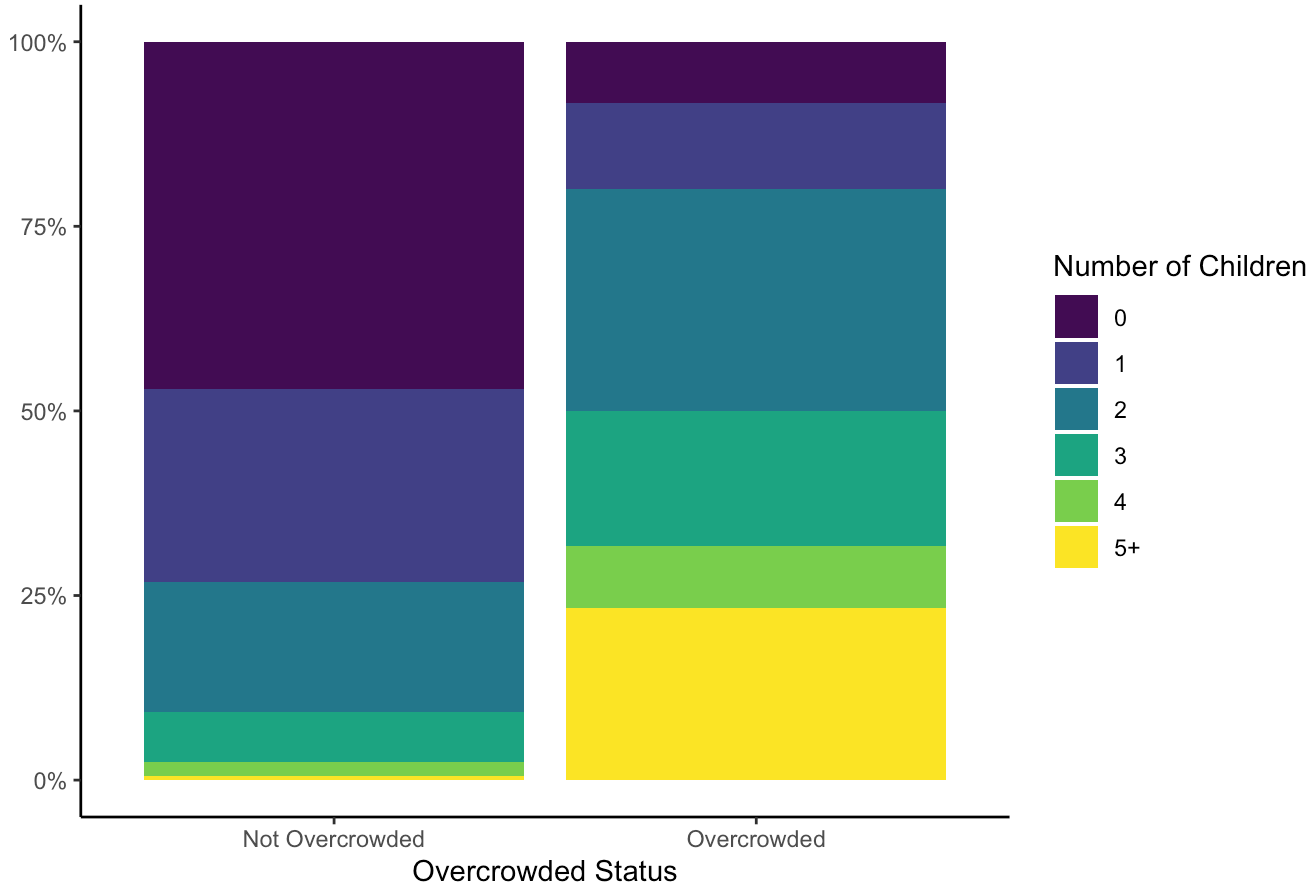

Figure 2 shows the relationship between the number of children a household has and if their household is overcrowded or not. We thought this was an important relationship between predictor variables, as the two statuses vary greatly for these categories. Looking at the coloring of the two bars we can see that there is a much greater prevalence of having between 0 and 2 children compared to having 3 or more for a household that is not overcrowded. On the contrary, for households that are overcrowded the number of children varies much more with all categories from 2 children and on being larger compared to households that aren't overcrowded. This relationship shows that overall, households are more likely to be overcrowded if they have more than 1 child.

Figure 3: Years of School by Poverty Status

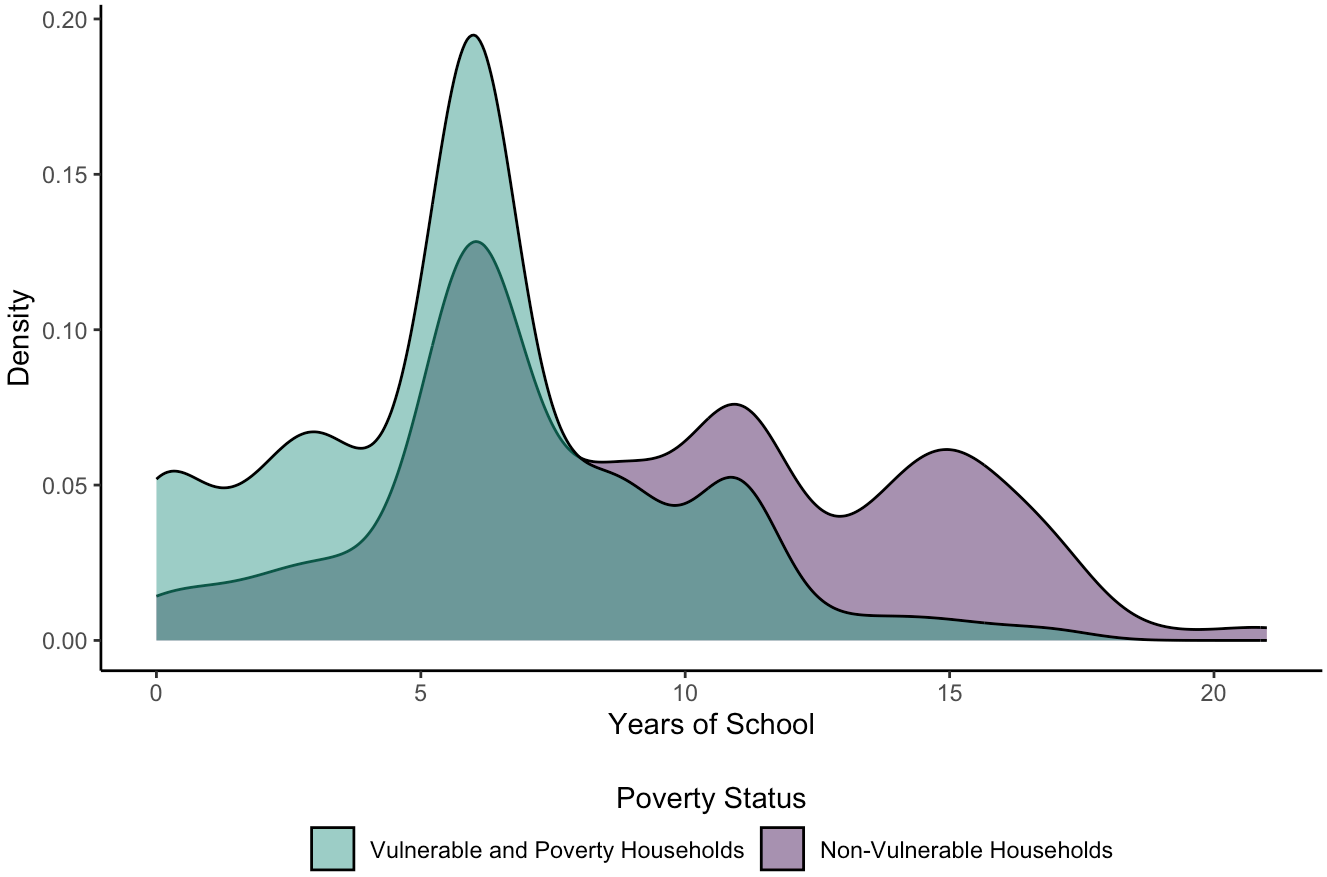

Figure 3 shows the relationship between how many years of school the head of the household has completed and their poverty status. Different from the original poverty levels, for this graph we split them into 2, with vulnerable and poverty households being levels 1-3, and non-vulnerable households being level 4. We thought this would be a good option for visualization purposes because many trends we looked at were very similar for levels 1-3, as those are the households that do have some sort of financial difficulty.

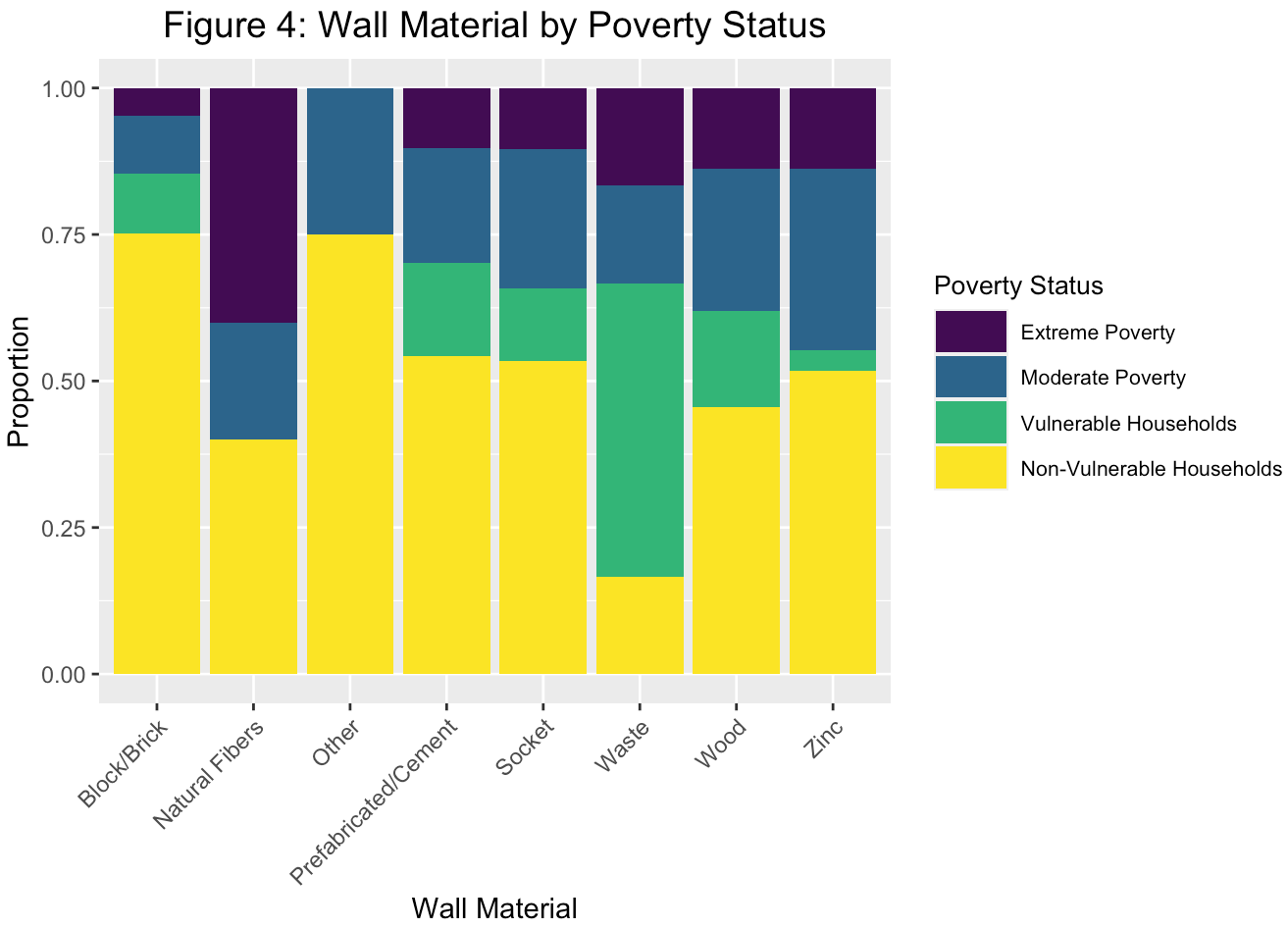

Figure 4 shows the proportion of each poverty status for different wall materials, showing that there is a difference in wall material based on the poverty status of an individual. The figure shows that there are no vulnerable households with natural fiber as a wall material. A large majority of vulnerable households have wall material that is considered waste. This figure also shows that those who are in the moderate poverty category have about equal proportions in each of the wall material categories. This means that individuals in the moderate poverty category have all different types of wall materials. The figure also shows that more individuals in the extreme poverty category have a wall material of natural fibers than other wall materials. Lastly, we see that individuals in the non-vulnerable household category have houses mainly made of blocks or bricks and other materials.

The Statistical Model

Description of the Model

We developed a prediction model to assess the poverty status of individuals based on various demographic factors. We chose a multinomial logistic regression model because poverty status has four possible categories: extreme poverty, moderate poverty, vulnerable households, and non-vulnerable households. To create the model, we included a comprehensive set of variables, excluding those directly related to the target variable and that we created from the other variables (such as different categorizations of the number of bedrooms). To select the most relevant variables, we explored various model selection techniques, including both forward and backward selection. After testing, we chose backward selection. This approach was more efficient and avoided issues we encountered with other methods, such as errors stemming from certain categories of floor materials in the testing data that were absent in the training data using forward selection. While both forward and backward selection produced identical evaluation metrics, backward selection took less time to run. Through backward selection, the

model reduced the number of variables from 41 to 17. The significant variables included factors such as whether the head of household has a bathroom, ownership of a tablet, the total number of males and females in the household, total household size, years of schooling, presence of a ceiling, disability status, the number of adults in the household, average education level, people per room, number of phones, age, type of toilet, roof condition, marital status, and geographic classification.

Results

When comparing the baseline model to the final model, the final model has far fewer predictors. This makes the model easier to interpret than the baseline. We can see this when comparing AIC scores. The baseline model has an AIC score of 3577.269 while the final model has an AIC score of 3425.673, suggesting that the final model balances goodness of fit and complexity better than the baseline. The baseline model does have a slightly lower residual deviance of 3079.269 compared to the final model with a residual deviance of 3275.673. This indicates that the baseline model fits slightly better which is often expected when including all possible predictors.

For the final model, extreme poverty was chosen as the baseline category. For a one-unit increase in a feature, the coefficient represents the log odds of being in another poverty class relative to the baseline. This interpretation gets easier to understand after exponentiating the coefficients. After applying exponentiation, the coefficients represent the likelihood of an instance being in a particular class compared to the baseline. For example, when looking at the feature has_bathroom, households with a bathroom are much more likely to be defined in classes 3 or 4 (vulnerable or non-vulnerable) than class 1 (extreme poverty) as the coefficients are both very high positive values before exponentiating (13.79 and 13.07). Typically these coefficients are not outside of the range of -1 to 1, but for the features describing if the household has a bathroom and the toilet type “none”, these coefficients are extremely large, indicating that not having a bathroom is a very telling indicator to if a person is in extreme poverty.

When looking at marital status as separated, we can see that the exponential coefficients are as follows: 1.037, 0.5235, and 0.4351 for categories 2, 3, and 4 (moderate poverty, vulnerable, and non-vulnerable). An individual whose marital status is separated is 3.76% more likely to be in moderate poverty than extreme poverty, 47.65% less likely to be in class vulnerable than extreme poverty, and 56.49% less likely to be in class non-vulnerable than extreme poverty. When looking at marital status as single, we can see that a person who is single is 78.01% more likely to be in moderate poverty than extreme poverty, 6.78% more likely to be in class vulnerable than extreme poverty, and 18.15% less likely to be in class non-vulnerable than extreme poverty. According to the final model coefficients, if a person is single or separated, they are more likely to be in class extreme poverty than non-vulnerable which is an interesting finding.

The final model using the training data had an overall accuracy of 69.5%. This model is doing well, considering that the model works to predict four classes. Digging deeper into the performance of the model, we found the sensitivity for each class to be 23.13%, 29.29%, 5.12%, and 95.31%. The sensitivity for vulnerable households is very low meaning that the model is not identifying people in this group well. We found that the specificity for each class was 97.73%, 93.27%, 97.99%, and 38.54%. The specificity for the non-vulnerable households is low. A low specificity for non-vulnerable households means that the model is not performing well at identifying observations that are not non-vulnerable households. These two evaluations lead us to believe that our model is over-categorizing people into non-vulnerable households. Our model has a weighted kappa value of 0.1945. This value is low but it is positive indicating that the model may be making many misclassifications between categories that are far apart such as predicting non-vulnerable when it should be extreme poverty.

The final model using the testing data achieved an overall accuracy of 66.1%. When analyzing the model’s performance, we found the sensitivity for each class to be 14.89%, 21.24%, 4.76%, and 93.85%. We found the specificity for each class to be 98.16%, 92.66%, 98.70%, and 28.69%. Again, we see that the model struggles in

identifying vulnerable households, and has trouble identifying households that are not non-vulnerable households. This is likely because non-vulnerable households make up about 66% of our data. Our model evaluating the testing data also does slightly worse than our model using the training data. This could be because the model is overfit to the training data. Overall, both models perform well considering the fact that we are trying to predict four classes instead of a binary class.

Figure 5: Confusion Matrix Heatmap of Testing Results

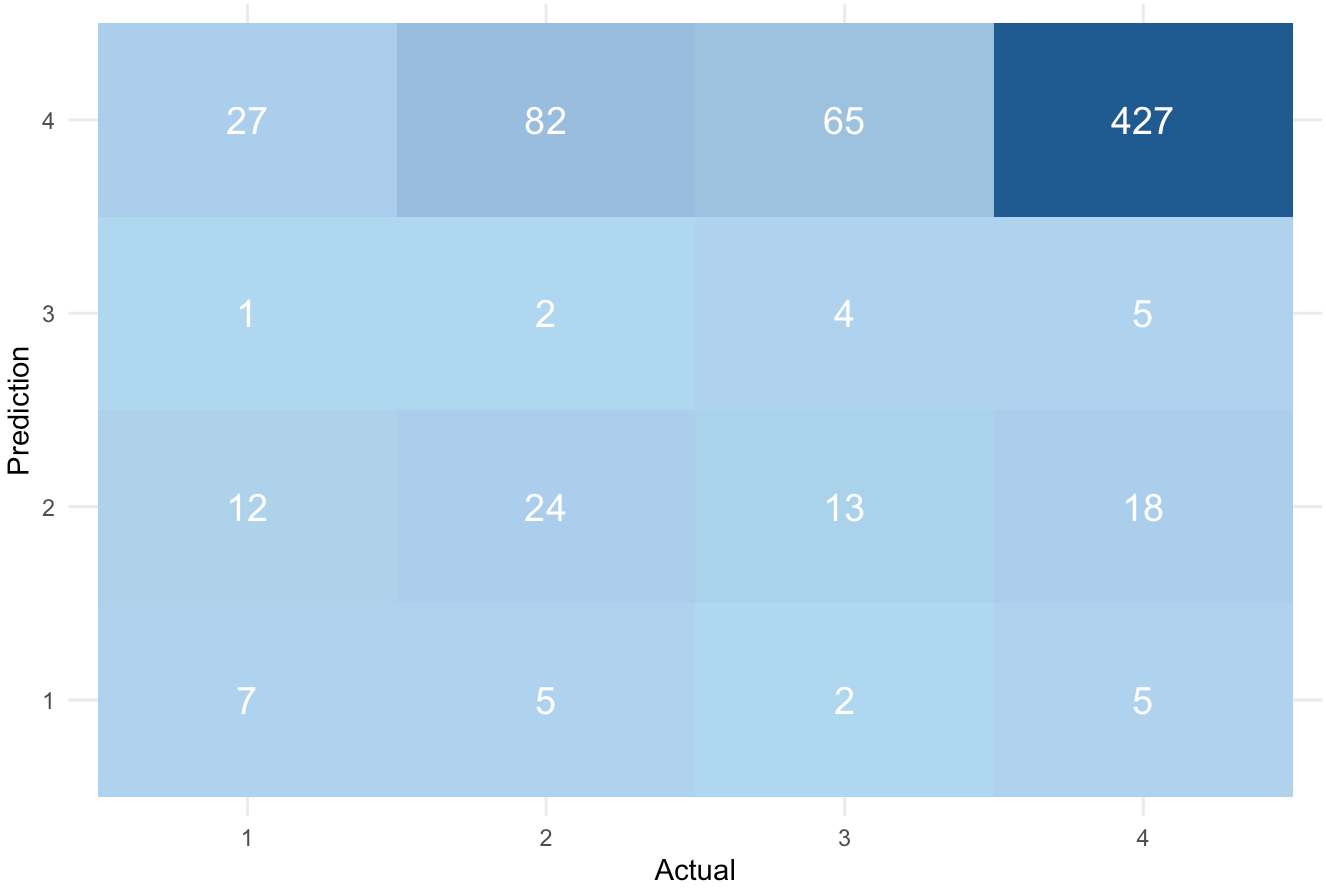

Figure 5 visualizes the results of the model on the testing data set. It shows the number of predictions for each group compared to the actual poverty placements. We can see that the majority of non-vulnerable groups (4) are being accurately predicted, along with this group being the largest group in the data. The smallest number of observations falls under vulnerable households (3), as the model struggles to predict this group accurately.

Our model can be a useful tool in identifying the classification of poverty for households in the region this data set was collected. This model could be a solid first step in identifying households that may need aid, but makes mistakes and should not be the only process used.