MA478 TEE

Tobias Hild

May 14, 2024

1 Data Exploration

Our training and validation data consists of 6002 observations of responses to a previous direct marketing campaign. We are trying to predict whether an individual donates in response to a direct marketing message as well as how much they will donate using variables including location, metrics related to their previous donations and the wealth of the donor's neighborhood. We first calculate the correlation of our variables, shown in Figure 1.

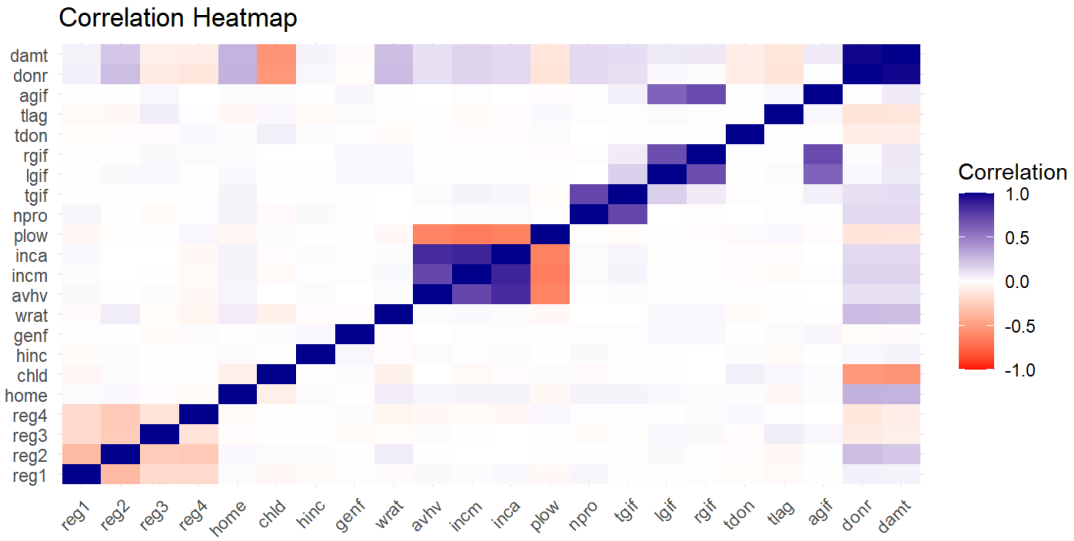

Figure 1: Heatmap of Variable Correlations

We find that the economic variables relating to the income in a potential donor's neighborhood are highly correlated. The variables most strongly correlated with donor amount are the number of children of the potential donor (-0.55), whether or not the donor is a homeowner (0.28), and the wealth rating of the donor's neighborhood (0.24).

We also found that several economic variables were correlated with the proportion of individuals that donated to the charity. For example in Figure 2 we can see that wealth rating, which is a metric that captures the wealth of a neighborhood using median family income and other population statistics, has a clear association which the proportion of donation responses received from individuals in that area. We should be cautious in this interpretation and any other donation response proportion results from our training dataset, since this dataset contains a higher proportion of responses than the actual data, and we are not sure whether this is a representative sample of responses.

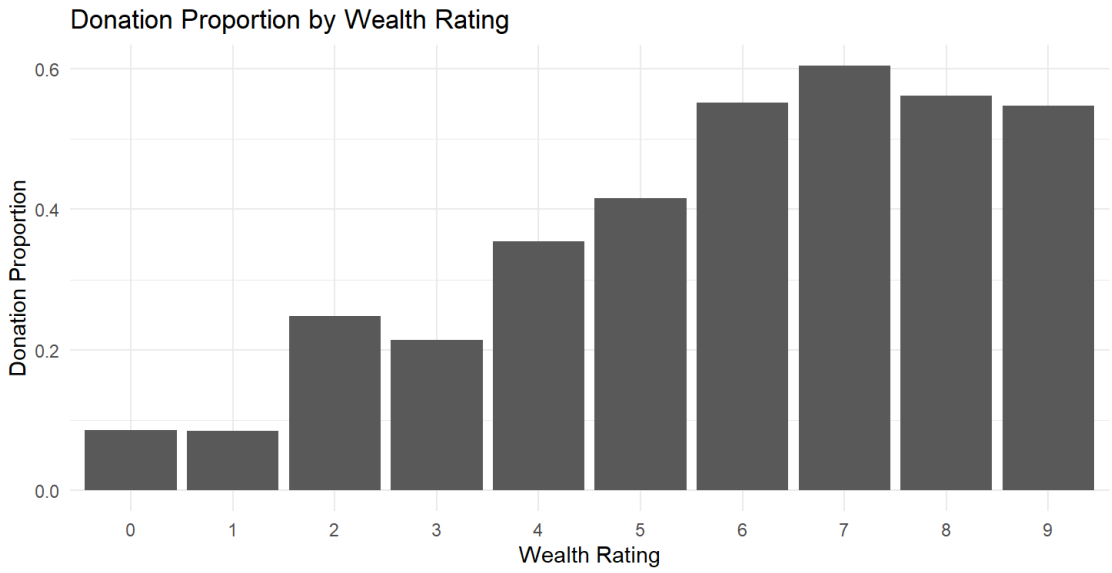

Figure 2: Wealth Rating

Before fitting models we standardized all explanatory variables. We encountered a number of variables which were strongly skewed (average home value, as well as income measurements); we transform these variables by taking the logarithm. We also develop a measure of the skew of the family income in an area by average family income and subtracting the median family income (for example a few very expensive houses in a neighborhood will result in a substantially higher average than median home value).

2 Modeling

We utilize a combination of two models to make a final prediction. The first model is a logistic regression which predicts the probability that an individual will make a donation. The second model predicts the size of the donation. This is a zero-inflated linear regression approach since it considers that there is some mechanism whereby the donation size is 0 (that the individual does not mail back a response) and sets an appropriate number of the predictions at 0.

Our first approach utilizes best subset selection for the linear component. That is for each possible number of variables we find the combination of explanatory variables the minimizes the BIC of the linear model on the validation set. We find that overall BIC is minimized with 10 variables and use these variables as explanatory variables for our linear model. We also found that using these variables minimized the mean squared error of the linear model on our validation dataset compared with other numbers of explanatory variables. We tried this approach

We also tried combining the binary and the quantitative components of the model into a single model using a zero-inflated model (we attempted both linear and poisson families). This model was trained on the full training dataset where the first approach's linear model only trained on the size of donations. While this result was promising on the training and validation datasets, it did not do well on the test dataset, we hypothesize because of the up-sampling of observations of actual donations in the training set. This approach might nevertheless be useful in future analysis on a dataset where the proportion of actual donations the same across training and test sets.

The final approach we tried was a generalized additive model. For this approach we utilized the variables which were selected by best subset selection above. However, in contrast to our finding with the linear model, where adding more variables than suggested by the best BIC did not result in a decrease of MSE, adding

additional variables to the GAM did decrease the MSE of the model on the validation dataset. We found our best results using 14 variables, the results of which are reported below. Splines were used for quantitative variables when the resulting model was significantly better than the same model using a linear predictor for that variable.

Parametric Coefficients

$$X = \begin{bmatrix} \text{Intercept} \\ \text{Region1} \\ \text{Region2} \\ \text{Region3} \\ \text{Region4} \\ \text{Homeowner} \\ \text{Number of Children} \\ \text{Household Income Category} \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} (14.24, 0.039) \\ (-0.045, 0.036) \\ (-0.08, 0.039) \\ (0.306, 0.037) \\ (0.635, 0.038) \\ (0.218, 0.055) \\ (-0.592, 0.034) \\ (0.504, 0.036) \end{bmatrix}$$

Significance of Smooth Terms

Variable	F-statistic
Median Neighborhood Income (1000$)	8.92
Neighborhood Low Income Proportion	12.07
Most recent gift amount ($)	16.03
Average gift amount ($)	12.45
Total gift amount ($)	8.96
Largest gift amount ($)	6.65
Time since last donation (months)	2.217

3 Analysis

We see from our parametic predictors, regions 3 and 4 are likely to have higher donations than the other three regions. We also find that being a homeowner is associated with generally higher donations. A greater number of children is associated with lower donation amounts. A great household income is associated with higher donation amounts. The number of promotions previously received has a slight negative association with donation amounts, although this variable was not significant in any model tested.

Of the smooth terms in our final generalized additive model the most significant were the terms for the size of the most recent gift, and the average gift. From our data exploration and other models, we believe that these are positive associations. We also see high significance for smooth economic variables such as median household income in the donor's neighborhood and the percentage of low income households in the donor's neighborhood. From our data exploration and previous models we believe that there is a positive association between median household income and gift size and a negative association between low income percentage and gift size.

Beyond the specific recommendations of our predictions we would suggest that the most profitable direct mailing opportunities exist in generally wealthy neighborhoods for individuals who have a history of giving larger gifts to the charity, but who have not received too many unaswered requests for money.