MA478 TEE

CDT Joshua Wong

May 14, 2024

1 Data Exploration

First, we looked at the distribution of our response variable, the donation amount given by each person who received a mailing. The distribution of the response variable can be seen in Figure 1. In this case, many zeros represent people who did not donate when sent a mailing and a separate distribution of people who did donate after the mailing. With such a separation in the response, we consider two different mechanisms that generate the zeros, whether or not someone donates, and the second mechanism represents a donation, how much a person donates. It is important to note that the second mechanism, which represents the amount of money donated given a donation, looks relatively normal.

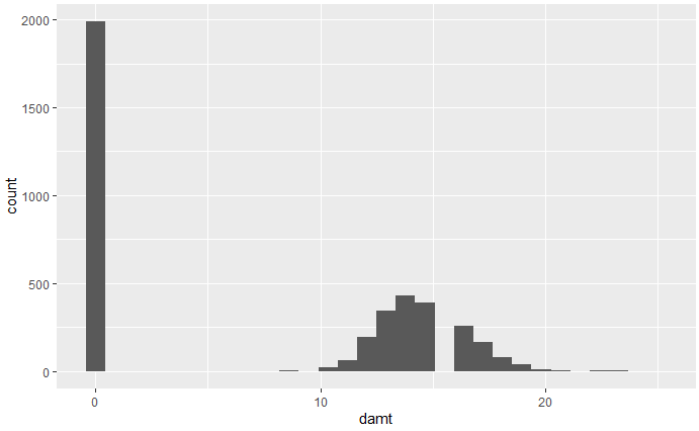

Figure 1: Histogram of Donations

Within our dataset we have 17 different variables that can help us predict our response variable. The information contained in these variables includes information about the individual, such as their sex, where they live by region, the number of children they have, their wealth, and whether or not they own a home. Other such variables look at the distribution of wealth within their neighborhood, including things like average home value, the median and average income of the neighborhood, or the amount of low-income people. Additionally, we include information about the donor's gift history, such as lifetime gifts, total dollar amount of gifts, time since the last gift, and time between gifts.

To look at the relationship between the variables at a high level we looked at the correlation matrix created between the numeric variables. Figure 2 displays the correlation matrix representing the correlation coefficients between each of our numerical variables. Notable is the high correlation between the number of children and the response variable of how much a person donates. Additionally, variables such as rgif and lgif representing the dollar amount of lifetime gifts and the largest gift have very high positive correlations with weather or not the person will give a donation given a mail email.

As for colinearity between the variables, we see very high colinearity between variables such as plow incm and inca, representing the percent low-income families and the average and median family income in the potential donor's neighborhood, respectively. Overall these relationships and the tyes of variables helped inform the variable selection for our eventual models.

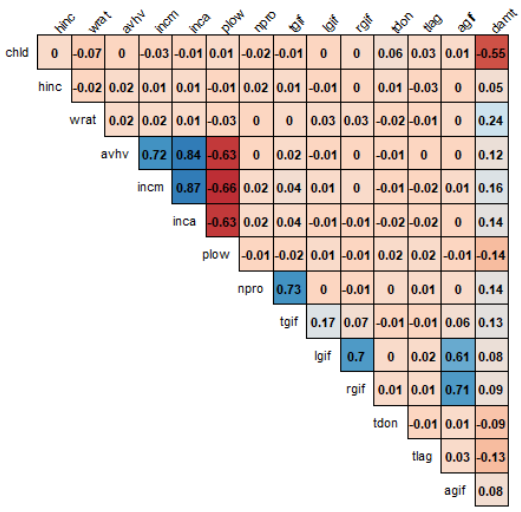

Figure 2: Correlation Matrix of Numerical Variables

2 Data Prepration

To prepare our data, we split the data into training and validation sets where all of our data exploration, initial analysis, and model building were done on our training set, and the testing was done on our validation set. We had 3984 training observations and 2018 validation observations that we tested our models on before deployment on the test dataset with the unknown predictor values. All of the metrics calculated, mean squared error, F1 score, and area under the curve (AUC) we calculated via how the model did on the validation dataset.

Additionally, in transforming our variables, we only used standardization, subtracting the mean and dividing by the standard deviation. This was to help with the predictability of the models so that all of the variables were weighted similarly on the same scale. We only standardized our numeric variables, not our categorical/binary variables. We also did not standardize our response variable.

3 Building Models

We considered three different candidate models to help predict our response variable. However, in all of the cases, we used a combination of two different models to help represent the two different mechanisms within the data. The first mechanism we saw from looking at our response variable was the large number of people not donating compared to the people who were. To predict who was and was not donating, we used logistic regression, as our data deciding whether or not someone was going to donate represented a binary response, coming from a Bernoulli distribution. Therefore, as we were tasked to develop generalized linear models for prediction, we chose a logistic regression for the binary component, with a Bernoulli random component and a logit link function.

On the other hand, to represent the second mechanism in our data, we looked at using a regular linear regression. We used a linear regression because, given that a person did donate, the distribution of donations was seen to be relatively normal. Therefore, as linear regression is a generalized linear model with an identity link function and a normal distribution for its random component, we considered it to be the best option for the second mechanism.

Our first model considered all of the variables in both mechanisms. In other words, the linear components of both our linear and logistic regression contained every variable provided to us to use, disregarding any potential collinearity. With this model, we found that our logistic regression had an F1 score of 0.84 and an area under the curve (AUC) of 0.914. For the linear component of the model we found a mean squared error of 1.88 and a standard error of 0.17.

Our next series of two models considered dealing with the colinearity within the data by removing specific covariates to see if the model performance increased. The first model we attempted simply used variable selection to determine which were the best based on the covariance matrix. In this

case, we use the variables of three of the regions, whether or not they were a homeowner, the number of children, wealth rating, the number of promotions they received, the total lifetime gifts, and the number of months since the last donation. We did not select any of the neighborhood variables initially, assuming that such information would be contained within the region. We also only selected one variable to represent the total amount donated one variable to represent the time aspect of the donations, and one variable, the wealth rating, to represent the overall wealth of the individual. In this case, our regression model was much worse, with a mean square error of 3.68 and a standard error of 0.26.

Our final model considered only disregarding the gender variable and the variable for one region, the fourth region from both models. We found that the gender variable was not significant given a nested chi-squared ANOVA test for both the linear regression and the logistic regression. We took out the fourth region variable, assuming the variation would be captured with the other regions. However, in this case, our linear regression still had a mean squared error of 2.11 and a standard error of 0.18. The F1 score and the AUC for the logistic regression were 0.84 and 0.91, respectively.

4 Model Selection

For our final model assessment and selection, we chose the fully loaded model as the best one out of our three, where both the logistic and linear regressions had the highest F1 score and AUC for the logic regression and the lowest means squared error for the linear regression. We found that removing covariates from the model only decreased its predictive capabilities. We think that for future work, instead of removing covariates, we should consider transforming and combining covariates to create new covariates that may help improve the predictive power of our models. Such transformations would be beneficial for this predictive task because the model would be able to consider more factors.