Predicting Donation Amount

CDT Jacob Hyatt

1 Data Exploration

Within the overall set of data, we have a large number of 0 dollars donated with the average amount donated at 7.21 dollars per household. Out of those who donate, the average donation is 14.50 dollars. Figure 2 shows the distribution of donation amounts from the full set of data. To explore the data further we looked at the distribution of all of the different variables. Some of the variables did not fit a nice Gaussian normal distribution. To change this we put them on a log scale. The variables that we changed are all of the variables relating to income (incm, inca,and avhv) and all of the variables concerning past gift amount (lgif, tgif, agif, and rgif). Changing these variables to their log counterparts increased every models performance.

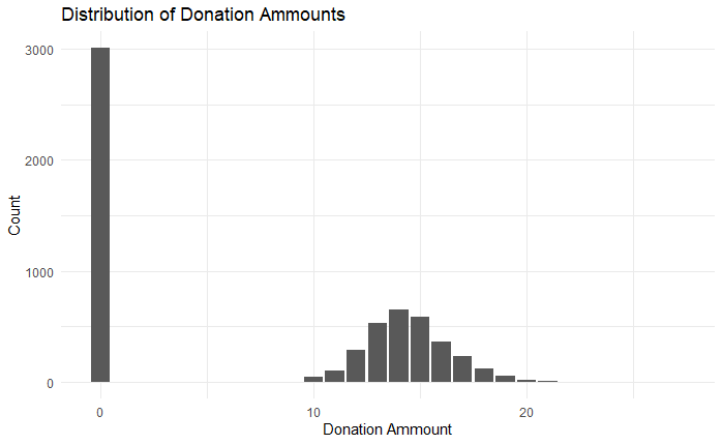

Figure 1: Distribution of Donation amounts within given data.

During our data exploration we created a correlation graph of the data. Average Home Value in potential donors neighborhood, Median Family income in potential donors neighborhood, and average family income in potential neighbors neighborhood all have a positive correlation with each other. While, percent categorized as "low income" in potential donors neighborhood has a negative correlation with the prior three variables. This observation makes sense, generally the higher the home values and income the less homes classified as low income are in the neighborhood. The bottom left side of the graph shows that the regions are all negatively correlated with each other. This shows that there are some differences in the regions. Figure 1 shows the correlation heat map.

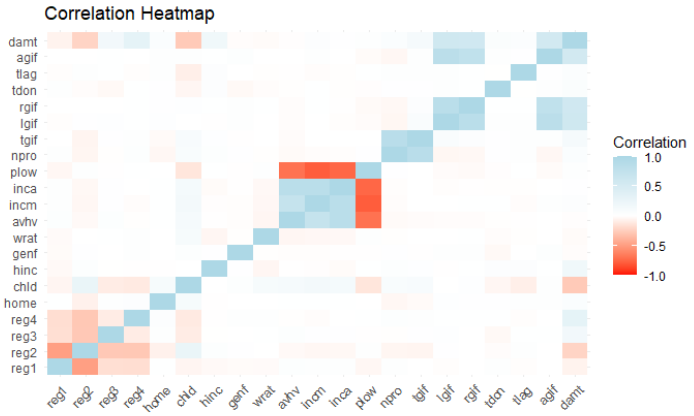

Figure 2: Correlation Heat Map

2 Model Building

Given our data exploration and research questions we believe the following models would be promising. A Gaussian generalized linear model with all of the variables from the dataset, a Gaussian generalized linear model with all of the significant variables from the dataset, and a Gaussian linear mixed model with the random component of percent categorized low income in potential donors neighborhood.

$$y_{ij} \sim \mathrm{N}(y_i)$$
$$\mu_i = \mathrm{N}(y_i)$$
$$\eta_{ij} = \beta_0 + \beta X$$

$$X = \begin{bmatrix} \text{Reg2} \\ \text{Reg3} \\ \text{Reg4} \\ \text{home} \\ \text{chld} \\ \text{hinc} \\ \text{genf} \\ \text{incm} \\ \text{tgif} \\ \text{lgif} \\ \text{rgif} \\ \text{tdon} \\ \text{agif} \\ \text{plow} \end{bmatrix}$$

Results for our simplified model are:

$$X = \begin{bmatrix} 1 \\ \text{Reg2} \\ \text{Reg3} \\ \text{Reg4} \\ \text{home} \\ \text{chld} \\ \text{hinc} \\ \text{genf} \\ \text{incm} \\ \text{tgif} \\ \text{lgif} \\ \text{rgif} \\ \text{tdon} \\ \text{agif} \\ \text{plow} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} 14.18091 \\ -0.05640 \\ 0.33532 \\ 0.66223 \\ 0.22695 \\ -0.59560 \\ 0.51062 \\ -0.05967 \\ 0.43912 \\ 0.20163 \\ 0.38867 \\ 0.47472 \\ 0.07110 \\ 0.38598 \\ 0.39674 \end{bmatrix}$$

dispersion parameter = 1.36

This model had an MSE of 1.558577 with a standard error of 0.1606044. This model had the lowest ACI of the tested models with an AIC of 6297.9. Being in Region 4, having more children, and being a male generally decrease the amount of money donated according to this model. All of the covariates in this model are significant.

However before even applying this model we use a classification simple binomial regression to find who even has a chance of donating. The starting model for this was a logistic regression that utilized all of the possible covariates. I added a squaring term to hinc and removed the avhv term. I added the squared hinc term because the distribution of hinc was off and the avhv had a p-value close to 1 so it was just causing undue noise in the data. The final confusion matrix can be seen in table 1.

Do not have time to add a table 1.

Would talk about and show the binomial regression model and talk about its stats. 1344.0 11652.5 c.valid chat.valid.log1 0 1 0 664 10 1 355 989

3 Analysis

Looking at the results of our models we can see that there is a clear winner with a lower AIC and a marginally higher MSE in Model 2. Model 2 is our simplest model as seen with the low AIC score, and on the test set it scored a .66787. This score is a large improvement on previously tried models. The results of all of the models can be seen in Table 2.

Model Name	**MSE**	**STD Error**	**AIC**
Model 1	1.556378	0.161215	6305.7
Model 2	1.558577	0.1606044	6297.9
Model 3	1.531064	0.161543	6365.438

Table 1: Values for Three Models

Model 3, the lmer model, had a lower MSE, but there was little improvement on the test set when applying the model. For decision makers, I would recommend focusing on the regional differences and neighborhood differences when using targeted advertising. I would also look at the binomial regression model to see if there is some improvement that can be made. The binomial regression has a lot of false positives. If the binomial regression model is improved, the targeted advertising can be further enhanced.