Predicting Donor Amounts- MA478 TEE

1. DATA EXPLORATION

The initial data set contains 8009 observations, including 24 columns surrounding around factors that contribute to donating to charities and if you donate or not. The main focus is surrounding around predicting the dollar amount of people who donated using generalized linear models. First, what was found was that there were only 2994 observations of donation. As seen in Figure 1, there is a mass majority of those who did not donate showing a lot of zeros in our data set for our response variable DAMT, dollars amount donated. Then when looking at figure 3, which shows the distribution for those donated it shows almost a normal distributed plot. Most people are 12-to-15-dollar range of for this variable. Then looking at correlation between our classification variable donor and predicting variable amount donated on figure 2, children amount have a negative correlation between both showing more children leads to people donating a lot or not donating at all. All the other explanatory variables do not show to much correlation besides a little with income related variables such as wrat, which is wealth rating and if you own a home. These conclusions are also shown in figure 4, which shows the boxplots of children amount relative to donated amount showing the less kids you have the more you donate. Then in figure 5, showing household income showing slight increases in dollar amounts when you are higher. These considerations will be factored in when deciding to chose variables for our classification model and model for prediction.

2. MODELS

I created three models for analysis for my prediction, which were Poisson Regression, Negative Binomial Regression, and then a Linear Regression Model. Then for my classification model I did a Binary Logistic regression which was used to test the set mailings to the most likely donors. The Binary Logistic Regression Model looked like this:

$$Log\left(\frac{1-p}{p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p$$

Where the response is the log odds of the observation being a donor. The coefficients for the model are two regions, homeowner, amount of children, household income, wealth rating, median and average income, plow, lifetime number of promotions, number of months since last donation, and number of months from first and last month. These covariates were picked from if they were significant in the big, saturated model. This was to reduce possible noise when trying to predict. Now for the prediction models the Poisson regression model looked like this:

$$Log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + .. + \beta_p X_{ip}$$

The Poisson regression model assumes that the mean and variance are equal, and the response represents the change in log counts for a one unit change in the predictor,

holding all other predictors constant. λ_i is the expected amount of donations for the i-th observation. The covariates here are two of the regions reg3, reg4, home owner, amount of children, dollar amount of recent gift, and average amount of dollar gift. These were the most significant variables in the saturated model and put into a single model. The next model I looked at was Negative Binomial to account for overdispersion in my model, this is the model:

$$Log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + .. + \beta_p X_{ip}$$

Additionally, the Negative Binomial regression model includes a parameter α to model the overdispersion: $Var(Y_i) = \mu_i + \alpha\mu_i^2$ where $Var(Y_i)$ is the variance amount of amount donated for the i-th ID. A value of $\alpha = 0$ would imply no overdispersion, reducing the model to Poisson Regression. This accounts for the limitations of Poisson. The covariates in this model are the same as Poisson to see the difference the results create. Lastly, I created a Linear Regression Model which looked like this:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + .. + \beta_p X_{ip}$$

Y_i is the dependent variable in this model for the i-th ID. The covariates in this model are two regions reg3, reg4, home owner, amount of children, household income, gender, average family income, plow, number of lifetime promotions, dollar amount of lifetime gifts to date, largest gift to date, recent gift, number of months since last donation and average dollar amount of gifts to date. These were chosen based on the significance of the saturated model for this model. The predictors quantify the change in the expected value of donor amount of Y_i associated with a one unit change in X_i.

3. ANALYSIS

When deciding what model was best for the prediction model for dollar amount between the three I based it on the Mean Square Error of each prediction of the model which is actual minus predicted squared and I only looked at the std error for the models.

MSE	Std Error	Model
140.6682	1.605547	Poisson Regression Model
140.6914	1.615674	Negative Binomial Model
1.867	0.169	Linear Regression Model

The Classification table for my binary logistic regression:

	0	1
0	709	18
1	310	981

Based on the MSE and Std Error the best model was the Linear Regression Model which showed extremely lower values than the other count regression models. Also, if you were going to pick between the count regressions Poisson and negative binomial it is best to pick Poisson for a simpler execution of counts. However, these models showed to be poor for this data. The model summary for Linear Regression model showed that reg3, reg4, home owner, income related variables, males, number of promotions all showed positive indications of increased dollar amounts donated. While, females, and if you had more children this also decreased the amount you donated. Below is the Model summary of more important variables:

Coefficients	Estimate	P-value
Intercept	14.17	< 0.001 ***
Reg4	0.67	< 0.001 ***
Home	0.25	< 0.001 ***
Amount of Children	-0.62	< 0.001 ***
Gender Female	-0.07	0.0224 *
Plow	0.21	< 0.001 ***
npro	0.13	0.00317 **

Higher income realates to donated more, males are more likely to donate more, and if you have more promotions, you also have a increased in donation amount. The residual standard error for this model is 1.281 with a significant F-statistic of 183.7. Then looking at the classification table the model did an overall job predicting if there were going not donated and did an alright job predicting if they were going to donate. The biggest conclusion of this analysis is that covariates such as amount of children and wealth indicators are good at alluding to if they are goin to donate or not and the amount. For future work it would be best to look at other models such as gamma regression which could help explaining the data a little better and predicting it better as well. The count regression models Poisson and Negative Binomial showed poor results in predicting and shouldn’t be looked at further.

APPENDIX

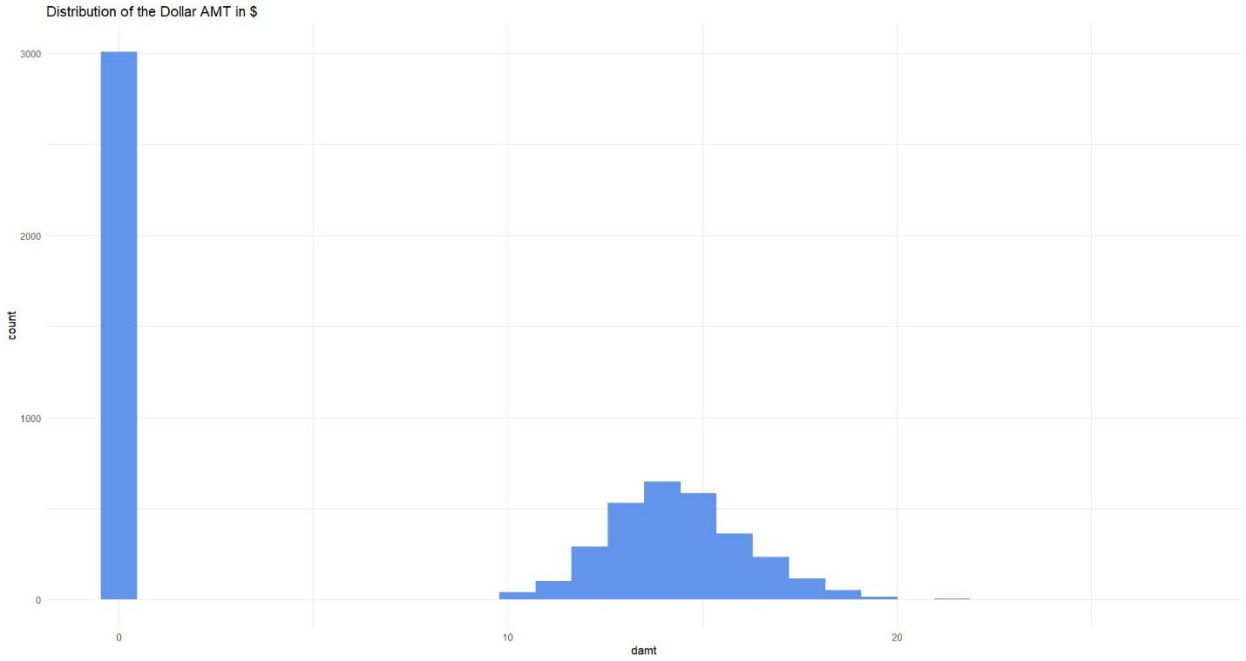

Figure 1: Histogram of the Dollar Amount Counts for Donating to Charity

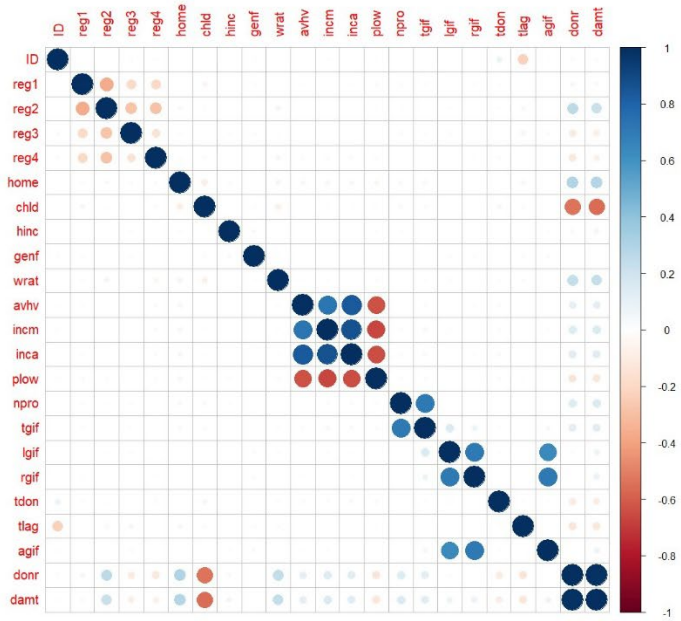

Figure 2: Correlation Figure of all the Variables in the Data Set

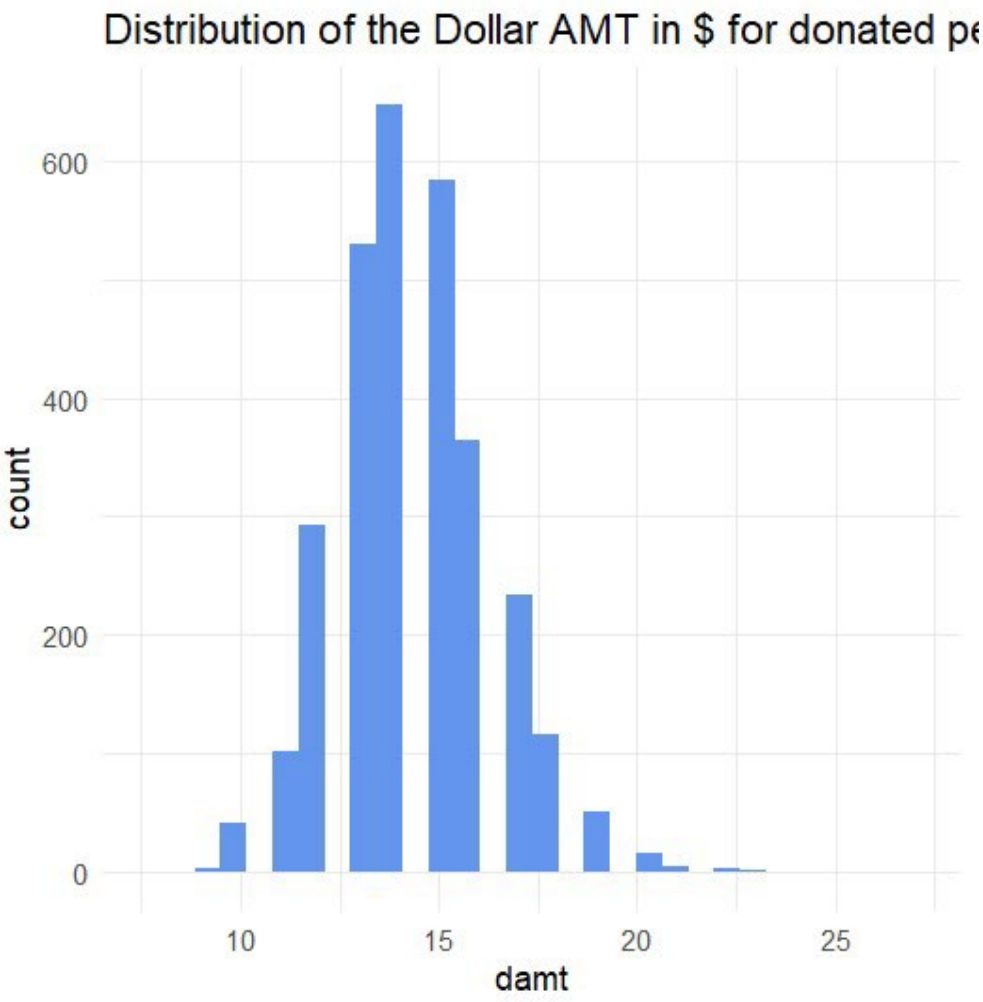

Figure 3: Histogram for people that donated, amount donated.

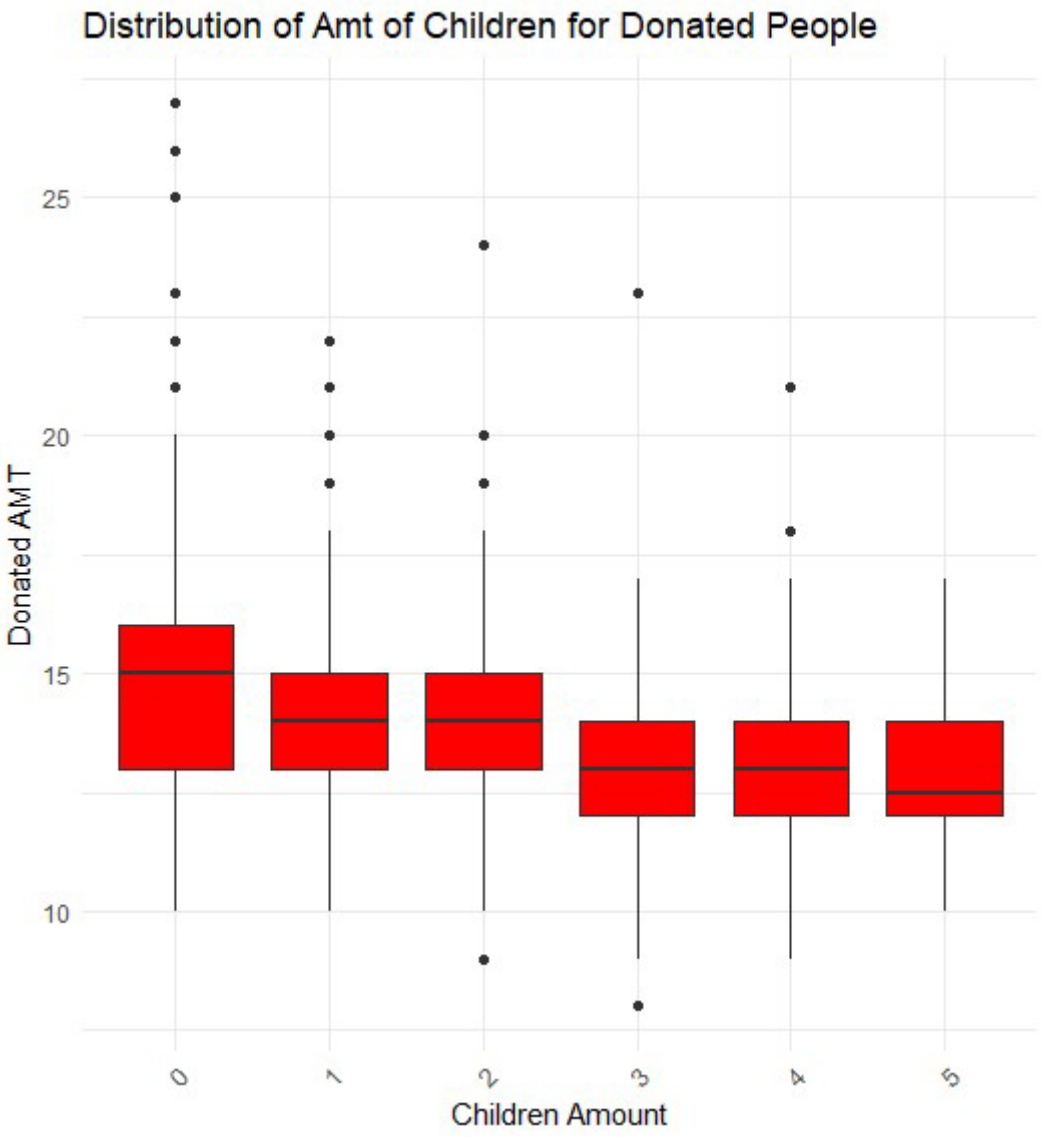

Figure 4: Boxplot of Children Amount with Amount donated for Donated people

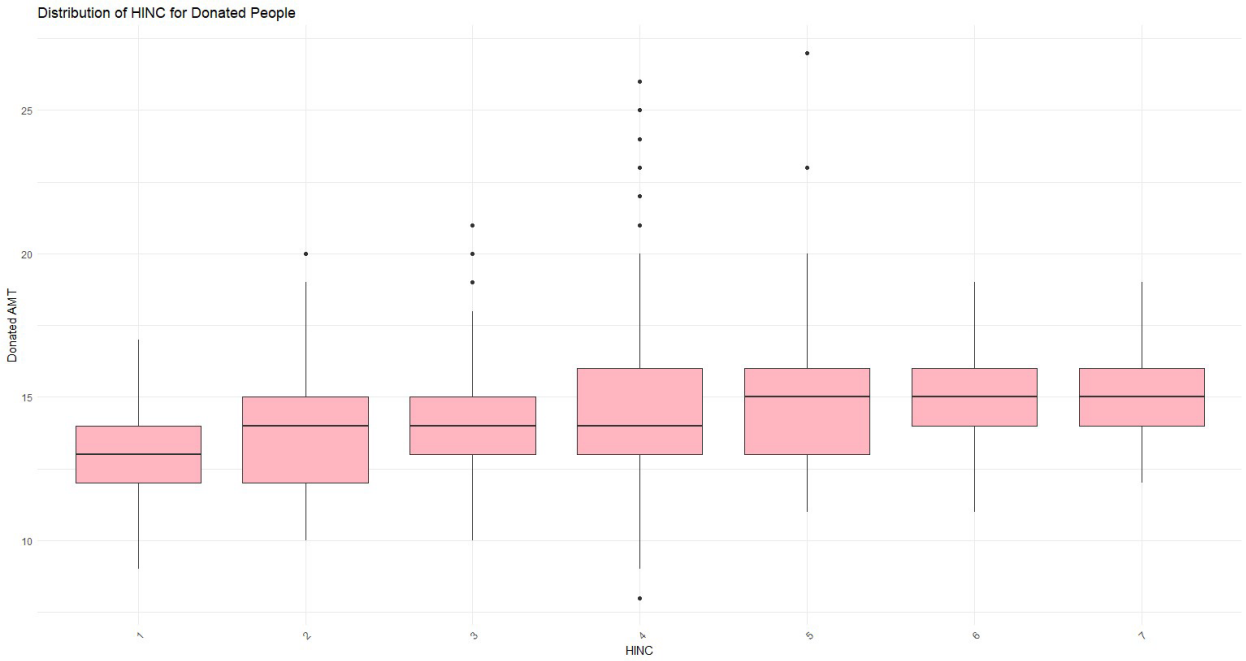

Figure 5: Boxplot of Household Income with Amount donated for Donated people