TEE Report

CDT Samin Kim

May 14, 2024

Data Exploration

The training dataset provided consists of 3984 observations. The data contains total 24 variables, which are economic factors of each donor. Out of 24 variables, the 'damt' variables indicate the amount of donation. The 'donr' variable indicates whether the person is a donor or not. Figure 1 shows the distribution of the 'damt' variable in the training set. The figure shows strong normal distribution but also shows excessive amount of 0's. This indicates that there are multiple people who did not donate at all. Figure 2 shows the distribution of 'damt' for only those who donated. The figure shows strong normal distribution for 'damt' variable, which suggests optimal model type for predicting 'damt'.

Figure 3 is the correlation heatmap of training dataset, showing the collinearity of across all variables in the dataset. Multiple variables showed collinearity such as 'avhv', 'incm', and 'inca'. All variables that showed collinearity were accounted when creating a model by having only a single variable of two variables present as predictor variable in the model.

Model

To predict the amount of donation, two types of models were used: Logistic regression model and the linear regression model. The logistic regression model was used to predict the 'donr' variable, which is the cause of excessive 0's in the dataset. By predicting the personal who is not likely to be a donor, we can predict the donation amount as 0. The logistic regression model was used because the 'donr' data consists of binary outcomes.

The linear regression model was used as Figure 2 shows strong normal distribution of the predicted variable, 'damt'. In this research, total three linear regression models were created and compared.

1. Linear regression model 1: containing predictor variables with statistical significance.
2. Linear regression model 2: mixed effects
3. Linear regression model 3: containing predictor variables showing strong linearity with 'damt'

Three types of models were created in the order that was listed above. For the first model, all variables in the dataset were included as a predictor variable, and the variables that did not show significant p-value were excluded. For the second model, the 'hinc' variable is a household income that contains total 7 categories. To examine the random effect of 'hinc' variable, the model with mixed effects was introduced. Lastly, each variable from the first model were inspected with a

visualization. Figure 4 and Figure 5 are examples of visual inspection. In this case, only 'rgit' was included as predictor variable since 'plow' does not show strong linearity with 'damt' variable.

Analysis

As a result, the first model showed the best performance out of three variables. Table 1 shows the AIC for each model, and the first model with only statistically significant variables showed the lowest AIC value with 6901.

Table 1: Summary of Model Evaluations with AIC

Model	AIC	MSE
Model 1	6901	2.12
Model 2	6962	2.12
Model 3	7001	2.24

The model 1 and model 2 showed very similar performance in terms of both AIC and MSE. This also indicates that the 'hinc' was significant in terms of predicting amount of donation, but random effects from 7 categories were not as significant when it was just treated as categorical variable.

The coefficients and p-value of each variable are shown in Table 2.

Table 2: Summary of Model 1

Variable	Coefficient	P-value
REG3	0.36079	$<$ 2e-16
REG4	0.64537	$<$ 2e-16
HOME	0.26792	3.37e-05
CHLD	-0.62366	$<$ 2e-16
HINC	0.50025	$<$ 2e-16
INCM	0.31560	$<$ 2e-16
PLOW	0.25879	5.95e-09
NPRO	0.18034	4.87e-09
RGIF	0.93820	$<$ 2e-16

The equation of the model 1 is the following (all variables are listed in Table 2:

$$\eta_i = \beta_0 + \beta_1 * (\text{REG3}_i) + \beta_2 * (\text{REG4}_i) + ... + \beta_9 * (\text{RGIF}_i)\epsilon_i$$

$$y_i \sim N(\mu_i, \sigma^2)$$

Where, i = donr, and $y_{ij} = \eta_i$ = Amount of donation made by donor i and j

From the model, we can tell that the 'RGIF' variable, which is the dollar mount of most recent gift has the highest impact on the amount of donation made, and 'PLOW' variable which is the percent categorized as "low income" in potential donor's neighborhood has the least impact on the amount of donation. In predicting

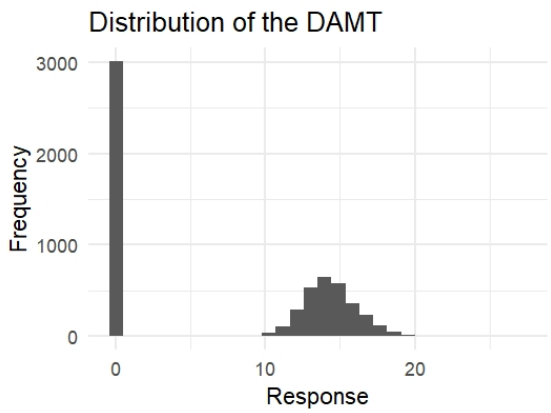

Figure 1: Histogram of DAMT in Training Set

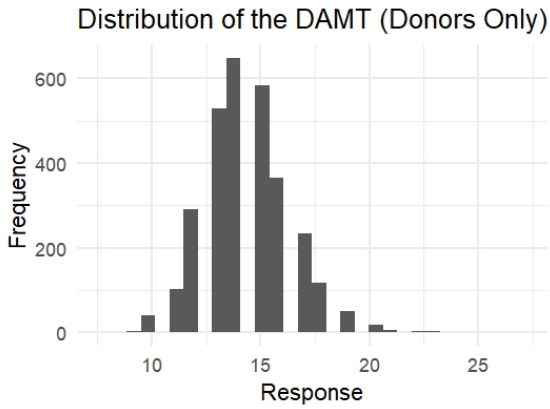

Figure 2: Distribution of DAMT (Donors Only)

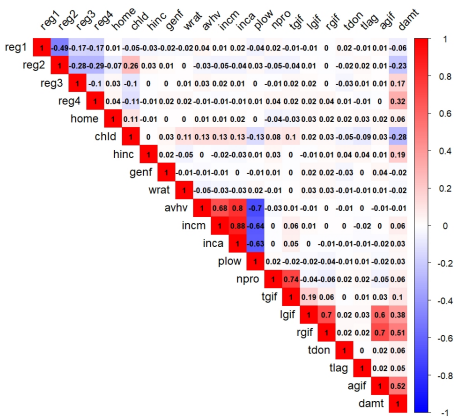

Figure 3: Correlation Heatmap

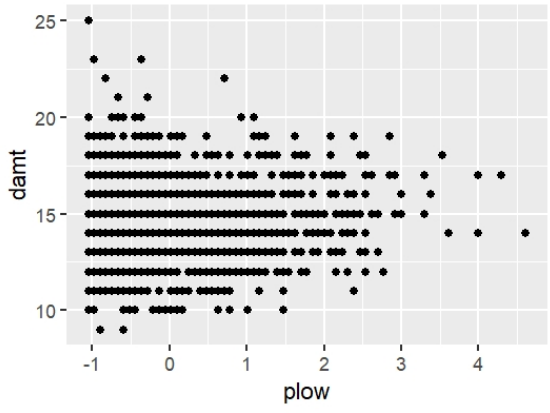

Figure 4: 'plow' vs 'damt'

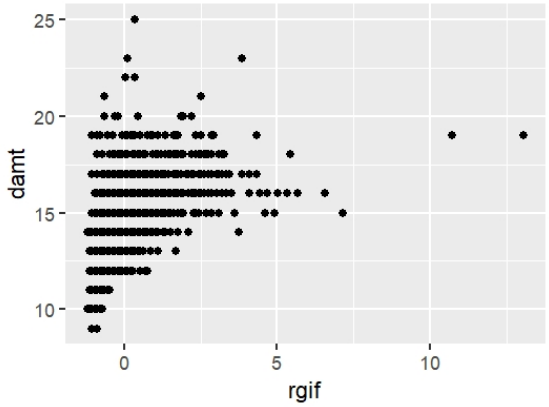

Figure 5: 'rgit' vs 'damt'