MA478 TEE

Evan Asuncion

May 2024

1 Data Exploration

The purpose of our study is primarily prediction. That is, our goal is to provide the charitable organization with an accurate model to predict potential donors as well as how much they will donate. To do so, we analyze a dataset of 6002 observations with 23 potential predictors. This dataset is relatively easy to work with as there are no missing values and data is relatively balanced. Before we create models, we must explore the nature of what we are predicting. Looking at potential donors, inherently, it is a binary random variable. That is, someone can be a potential donor or not. When we look at the distribution of the donor variable, we see that the data is balanced with approximately 3,000 donors and non-donors.

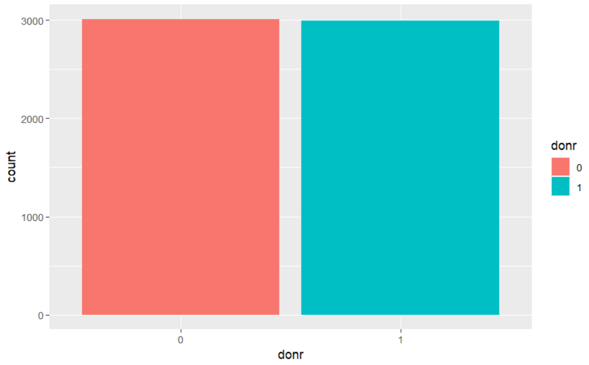

Figure 1: Observed Distribution of Donors or not

The number of donors versus non-donors being balanced will help in our analysis. The distribution also stregnthens our assumption that the random variable follows Bernoulli distribution. Looking at the other response variable in question, amount donated, we see that we should implement linear regression to predict these values.

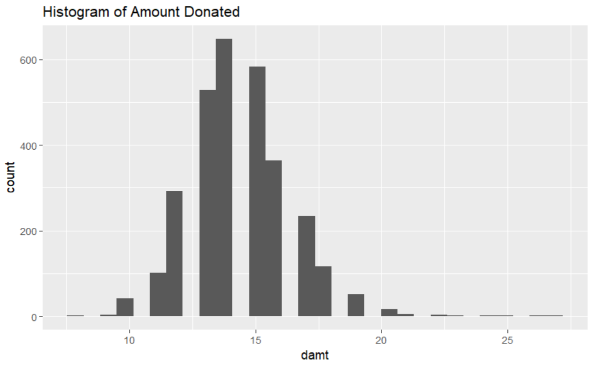

Figure 2: Observed Distribution of Amount Donated

With a mean of $14.50 donated on average and a standard deviation 1.98, this distribution is roughly normal. The shape of the variable's distribution combined with the quantitativ nature of the amount gifted, makes generated GLMs with a gaussian family and the identity link function to be the right choice.

The 23 potential predictors are primarily socio-economic in nature–income, region, and if the potential owner owns a home are some examples. As such, some our closely related. Figure 3, helps us identify any variables that are potentially collinear while also suggesting variables that we should add to our initial models.

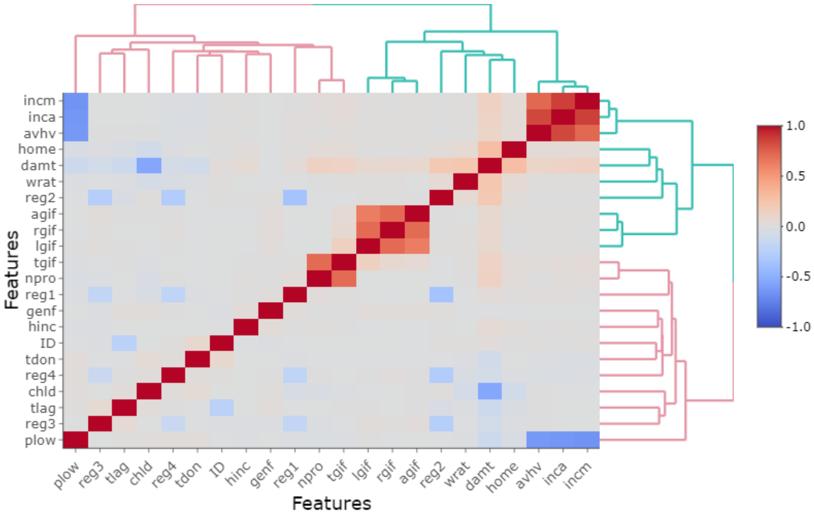

Figure 3: Correlation Heatmap of Predictors and Quantitative Response

From the figure, we see that number of children, being a homeowner, and other factors relating to wealth have a somewhat significant sample correlation coefficient when compared to amount donated. Furthermore, we see that variables with similar descriptions–averages, medians, mins, and maxes–exhibit potential collinearity. As such, we will only include one of these measures for each class of similar variables.

2 Model

2.1 Variable Selection

To select variables to include in the models, we implement best subset selection. The plot resulting from best subset selection did not fully plateau at what it showed, so we chose the eight most significant variables from the selection process. In the future, we could implement a more efficient method such as Lasso or Ridge regression in variable selection.

2.2 Linear Regression

From our best subset selection, the first model we look at is the following:

$$\begin{aligned} i &= observation \\ y_i &\sim N(0, \sigma^2) \\ \eta &= y \\ y &= X\beta \end{aligned} \tag{1}$$

Where X is a vector containing observation i's reg1, reg2, reg4, chld, hinc, rgif, agif, and damt values. However, this model failed the goodness of fit test.

We assume there to be a difference between genders when it comes to income and number of children, so the second model we consider includes an interaction terms between both gender, income, and number of children. Again, we do not find this model to be better than the saturated model.

As such, the third model we consider is the initial model but with a "probit" link. This different link function assumes a latent process. However, it yielded a lower AIC than the probit link function, so we do not consider it. It is also less interpretable.

2.3 Logistic Regression

The first model we considered, like the initial model, was a basic Bernoulli GLM with a logit link and linear predictor coefficients relating to variables from the best subset variable selection process.

To verify these models to be better, we conduct Chi-square goodness of fit tests to see if these models are worse than the saturated model.

Since there is a difference in the number of donors according to region, and each region could have its own random uniqueness, we consider a GLMM as our next model. However, this model yielded a lower AIC than a nested model one step less complex.

As none of the prior models we consider beat the saturated model, we present a GLM centered aorund all the predictors in the dataset as my final model. Again, this is not ideal and should not be implemented in prediction.

3 Analysis

If I had more time I would comment on how I failed to generate predictive models. All of the MAE's that I derived were terrible. However, I could comment on some inferential findigs. Such as the significant effect of number of children on amount given.