Your Paper

You

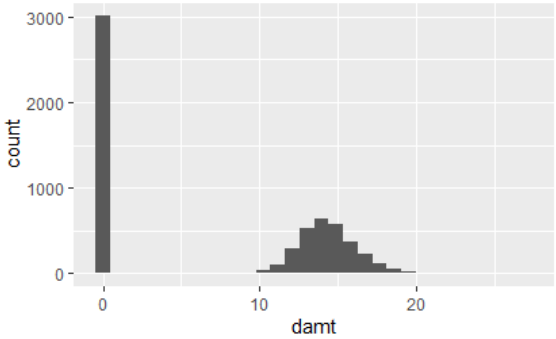
May 14, 2024

1 Introduction

Propose a model to use for improving the cost-effectiveness of a charitable organization's direct marketing campaign. Currently at an overall response rate of 10%. The average donation from those who respond is $14.50. With the current response rate, donations received, and cost of sending, the organization is currently losing money. To maximize the net profit, a cutoff is recommended for determining based on information from a person, if they will donate and how much can be expected from them.

2 Data Exploration

First for exploring the dataset given, which consists of 3984 observations to train on, 2018 for validation and 2007 for testing. The initial exploration of the training data showed a large population of 0's for donations that were observed throughout the data.



Count of donation values.

While there are a large number of 0's, this is to be expected, as only 10% of people are observed responding to donate. From here, looking at the data to see who responded, the information shown in the variables INCM, INCA, AVHV, have very large values that are skewing the predictions of the model. Transforming to take the log of those values was an approach I tried and was successful with in increasing the prediction ability of the model during testing.
* I chose variables off the list that I thought would show co-linearity, I was able to confirm these predictions viewing the pairs plots of select variables together. When looking at all variables together at the same time, no patterns were able to be seen as the information was too small. After noticing which variables were

3 Models

$y_i = \beta_o + \beta_1 x_i + \epsilon_i$ $i = observationalunit, y_i = responsevariable, x_i = covariate/explanatory\_variable, \epsilon_i \sim N(0, \sigma)$ Binary logistic regression was used to fit that model and linear regression was used to fit the model that provided estimates on the amount of money each predicted donor would likely contribute.

3.1 Linear Regression Model 1

$model1 < -glm(damt\ reg1 + reg2 + home * chld + I(hinc^2) + genf + wrat + incm + plow + npro + tgif + tdon + tlag^2, data.train.std.c)$
* This model was produced before the parameters were transformed, additionally, I took out reg3 and reg4 because when looking at the summary data, they showed as insignificant. However, when adding them back into the model, the predictive power was increased and so were savings.

3.2 Linear Regression Model 2

Model 2 was built using binary logistic regression, the parameters of this model, $model2 < -glm(damt\, reg1 + reg2 + reg3 + reg4 + home + chld + hinc + I(hinc^2) + genf + wrat + avhv + incm + inca + plow + npro + tgif + lgif + rgif + tdon + tlag + agif, data.train.std.c)$
* estimated savings of $3998.58. This was the best result of any model produced. The only difference between this model and model 3 was squaring the hinc variable. This did make a difference, as model 2 performed better, but not by very much.

3.3 Linear Regression Model 3

The third model is similar to the 2nd, in that it also $model3 < -lm(damt\ reg1 + reg2 + reg3 + reg4 + home + chld + hinc + genf + wrat + avhv + incm + inca + plow + npro + tgif + lgif + rgif + tdon + tlag + agif, data.train.std.c)$
* This model differed from the other models in that it still contained all parameters, similar to model 2, but did not have the transformation on variable hinc. The variables inca, incm and avhv, I still took the log of them, but the model performed worse at predicting when read through the kaggle submission. Additionally, the estimated savings for this model were less than 3998.58 given in model 2. Model 2 also gave a better MSE

4 Analysis

The best-performing model for predicting who will donate was model 2.