UNITED STATES MILITARY ACADEMY

MA478 TEE

MA478: GENERALIZED LINEAR MODELS

SECTION H2

COL NICHOLAS CLARK

By

CADET ISABELLA PALCHAK ‘24, CO D1

WEST POINT, NEW YORK

14 MAY 2024

_______ I CERTIFY THAT I HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT I USED TO COMPLETE THIS ASSIGNMENT AND THAT I ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

_______ I CERTIFY THAT I DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTANCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURE: ______________________________________________________________

# MA478 TEE

CDT Isabella Palchak

May 14, 2024

**Abstract**

This report summarizes the work produced during our MA478 TEE. To complete the TEE we were tasked to create two generalized linear models (GLMs). Our client is a charitable organization that seeks to improve the cost effectiveness of its marketing campaigns to previous donors. This task comes with two questions: Can statistics on previous donors be utilized to predict who will donate in the future? Can statistics on previous donors predict the donation amount?

## 1 Introduction

The provided data was split into a training set, validation set, and test set. The training set consisted of 3984 observations, the validation set consisted of 2018 observations, and the test set consisted of 2007 observations. The data had 24 variables that were potential predictors of future donors and future donation amounts. Again the two tasks were as follows:

- Develop a GLM using data from the most recent campaign that can effectively capture likely donors so that the expected net profit is maximized.
- Develop a GLM to predict donation amounts for donors - the data for this will consist of the records for donors only.

Variables describing donors consisted of geographic variables, socioeconomic variables, and statistics on past donations. Potential covariates included region that the donor lives in, whether or not a donor is a homeowner, the number of childern the donor has, and whether or not the donor is male or female. Variables also capture measures donor's income and the wealth of the donor's neighborhood. Additionally, variables capture the size and amount of donations received.

## 2 Data Exploration

For data exploration we first looked at the data we wanted to classify. Whether or not an individual will be classified as a donor or non-donor. You can see in *Figure 1* in Appendix A that there are a lot more non-donors than there are donors.

Then we investigated the potential predictors. We looked at all the variables and determined whether or not they would be useful in predicting the binary outcome of donor or not donor. We grouped the training data by the outcome (donor or non-donor) and looked at the average of each variable. For example, the average of the variable INCA (average family income in the potential donor's neighborhood) was much higher in for donors ($59,900) than it was for non-donors ($52,700). Similarly, WRAT (a measurement of donor's wealth) also appears to be higher for donors (7.58) than it is for non-donors (6.47). We eliminated a lot of poor predictors of whether or not an individual is a likely donor using this method. I eliminated variables such as AGIF (average donation amount), LGIF (largest donation amount), AVHV (average homevalue in the donor's neighborhood), and RGIF (most recent donation amount). There was negligible different between the average of these variables for donors compared to the non-donors. This helped us to determine what variables to include in our first GLM that classifies individuals as donors or non-donors.

Finally *Figure 1* in Appendix A shows that there is a significant number of zeros in the observations. The count plot shows on the x-axis the donation amount and the frequency on the y-axis. There are significantly more non-donors than there are donors. From this information it seems we may want to consider a zero-inflated model to account for the significantly large count of zero as the donation amount. This information is useful for the second objective–building a GLM to predict the donation amount that the donor will gift.

## 3 Data Preparation

To prepare the data, we transformed the RGIF (amount of most recent donation), LGIF (amount of largest donation), TGIF (total amount of all donations), AVHV (average home value in donor's neighborhood), and INCM (average family income value in donor's neighborhood) by taking the log of each variable. We did this because the distribution of these potential predictors are skewed and some GLMs are sensitive to skewed predictors.

## 4 Build Models

I will present models that accomplish both of the assigned tasks. First I will present logistic regression models with a logit link that will be used to classify individuals as either a donor or non-donor. Then I will present linear regression model that will be used to predict the dollar amount that future donors will make.

$$\hat{log}(\mu) = \beta_0 + \beta_1 * TLAG + \beta_2 * NPRO + \beta_3 * PLOW + \beta_4 * WRAT$$

| Model 1 Logistic Regression | | |
|---|---|---|
| Predictor | Fixed Effect | P-value |
| $\beta_0$ | -0.02212 | 0.515 |
| $\beta_1$ | -0.32114 | $< 2e - 16$ |
| $\beta_2$ | 0.30632 | $< 2e - 16$ |
| $\beta_3$ | -0.29620 | $< 2e - 16$ |
| $\beta_4$ | 0.56849 | $< 2e - 16$ |

$$Donation\_Amount = \beta_0 + \beta_1 * TLAG + \beta_2 + NPRO + \beta_3 * PLOW + \beta_4 WRAT$$

| Model 1 Linear Regression | | |
|---|---|---|
| Predictor | Coefficient | P-value |
| $\beta_0$ | 14.52411 | $< 2e - 16$ |
| $\beta_1$ | 0.11007 | 0.0300 |
| $\beta_2$ | 0.10555 | 0.0146 |
| $\beta_3$ | 0.06197 | 0.1969 |
| $\beta_4$ | -0.06347 | 0.3081 |

This logistic regression model had and AIC of 5032.1 and reported a maximized profit of $10,518 and 1940.0 mailings send to potential donors. All of the predictors in this model are statistically significant with a p-value of less than $2e - 16$. There appears to be a negative relationship between donors and TLAG as well as PLOW. After adjusting for oversampling adjustment by calculating number of mailings for test set, ourlinear regression model suggests that the organization mail to the 1474 highest posterior probabilities. The model suggests that the donors will contribute a total of $18453.93 to the charity.

Model 2 is shown below. Model 1 is nested within this model. This means that the predictors in Model 1 are captured in model 2. It includes additional predictors of HOME and INCA which indicate whether or not a donor is a homeowner and the average household income of their neighborhood:

$$\hat{log}(\mu) = \beta_0 + \beta_1 * TLAG + \beta_2 * NPRO + \beta_3 * PLOW + \beta_4 * WRAT + \beta_5 * HOME + \beta_6 * INCA$$

$$Donation\_Amount = \beta_0 + \beta_1 * TLAG + \beta_2 + NPRO + \beta_3 * PLOW + \beta_4 * WRAT + \beta_5 * HOME + \beta_6 * INCA$$

The logistic regression indicated that a total of 1940 mailings should be sent and a maximized profit of $10518.5 would be returned. The linear regression included two new covariates: HOME and INCA, which had the estimates of 0.23721 and 0.15407 and p-values of 0.00878 and 0.00263 respectively. had a mean squared error of 4.267192 and a standard error of 0.29022

The third model is shown below:

$$\hat{log}(\mu) = \beta_0 + \beta_1 * TLAG + \beta_2 * NPRO + \beta_3 * PLOW + \beta_4 * WRAT + \beta_5 * HOME + \beta_6 * INCA$$

$$Donation\_Amount = \beta_0 + \beta_1 * TLAG + \beta_2 + NPRO + \beta_3 * PLOW + \beta_4 * WRAT + \beta_5 * HOME + \beta_6 * INCA$$

The logistic regression suggested that there will need to be 1829 mailings sent and a profit of $10682.5 returned. Based on this model, the organization should send mail to the 385 highest postierior probabilities. The linear regression had a mean squared error of 3.16 and a standard deviation of 0.2439, this model predicted estimated profits at $4782.83 in total donations.

## 5 Select Models

We choose the first model because it is the least complex model with the lowest mean squared error. The second model has more complexity but does not significantly improve the poroft returned by the donors or reduce the number of mailings that need to be sent. It seems that TLAG, NPRO, PLOW, WRAT, HOME, and INCA all are predictors for whether or not an individual will be a future donor but are not good predictors for the amount of money that a donor will donate.

# 6 Appendix A

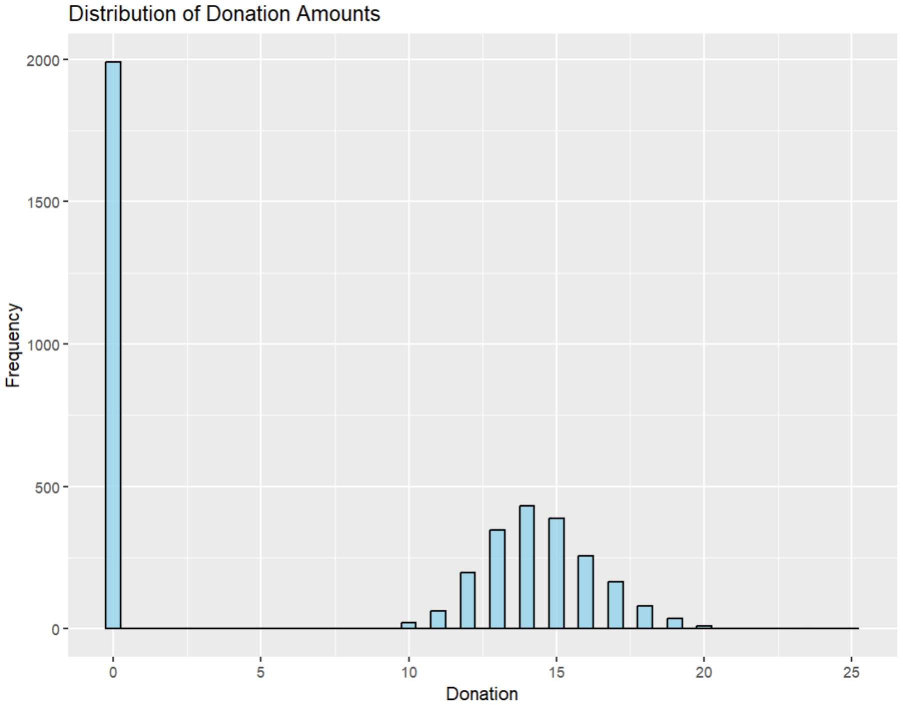

Figure 1: Distribution of the Donation Amounts for Previous Donors