Final Tee

Aimee Rohan Ramirez

May 2024

1 Introduction

Understanding the dynamics of donor behavior is crucial for optimizing the efficiency of fundraising campaigns for charitable organizations. In this study, we aim to develop generalized linear models (GLMs) to enhance the cost-effectiveness of direct marketing efforts targeting previous donors.

The dataset provided consists of mailing records from a recent campaign conducted by the charitable organization. With an overall response rate of 10%, and an average donation of $14.50 from responders, our objective is to identify predictors that accurately distinguish likely donors from non-donors. By doing so, we seek to maximize the expected net profit per mailing, considering the cost of production and distribution.

Additionally, we endeavor to build a GLM capable of predicting the expected gift amounts from donors exclusively. This model will enable us to tailor fundraising strategies more effectively, optimizing resource allocation and campaign outcomes. Through rigorous data exploration, preparation, and model building, we aim to provide actionable insights and predictive tools to empower decision-making processes in future fundraising endeavors.

2 Data Exploration

How do given characteristics about previous donors impact whether or not they donate or how much they donate? In this project we were given 3984 training observations, 2018 validation observations, and 2007 test observations.

In conducting a univariate analysis on region, it's evident that region two comprises the largest proportion of individuals on the mailing list, as depicted in Figure 1. This suggests that region two might have a higher representation in the dataset compared to other regions.

Furthermore, when examining the distribution of the number of children per individual, the histogram in Figure 2) reveals that the majority of individuals have either no children or between one to two children. This distribution provides insight into the family demographics of the dataset, indicating that households with fewer children are more prevalent. Upon examining the multicollinearity matrix as well, a noteworthy finding surfaces: the presence of children demonstrates a negative correlation with both donation likelihood and donation amount (Table 2,Table 1). This indicates that individuals with children tend to exhibit lower propensities for donation and contribute smaller amounts.

In scrutinizing the distribution of donors and non-donors within the dataset, it's crucial to recognize the balanced distribution observed, as depicted in Figure 3. However, a notable observation emerges when considering donor distribution concerning gender. As illustrated in Figure 4, females not only constitute a larger portion of the dataset but also exhibit higher counts in terms of donor representation.

3 Data Preparation

I did not classify the variables that have multiple levels as factors and kept them as quantitative because it helped maintain model simplicity and conciseness, particularly when dealing with a large number of levels.

In steps done for me, the predictor variables were standardized to have a mean of zero and a standard deviation of one to ensure uniform scales across features. Standardization was performed separately for the training, validation, and test datasets to prevent data leakage and maintain consistency. The standardized predictor variables were then used to create standardized training, validation, and test datasets, where the 'donr' variable is included for classification and the 'damt' variable is included for prediction when 'donr' equals one. This process ensures that the statistical models are trained, validated, and tested on standardized data, facilitating model interpretation and comparison across datasets.

4 Build Models

4.1 Binary Logistic regression and multiple linear regression

Logistic Regression 1

This is a logistic regression model using the logit link as it is very interpretable. The predictors were carefully chosen based on the results of data exploration, specifically selecting variables with significant associations with donor classification, as shown in Table 1. In terms of interpretation, each coefficient represents the change in the log-odds of the event (being a donot) happening for a one-unit change in the predictor, holding other variables constant. The signs of the coefficients indicate the direction of the effect on the log-odds.

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) =& \beta_0 + \beta_1 \cdot \text{region2} + \beta_2 \cdot \text{home} + \beta_3 \cdot \text{children} + \beta_4 \cdot \text{household\_income} + \beta_5 \cdot I(household_i ncome^2) + \\ & \beta_6 \cdot \text{wealth\_rating} + \beta_7 \cdot \text{income} + \beta_8 \cdot \%\text{low\_income} + \beta_9 \cdot \text{number\_promotions}\end{aligned} \tag{1}$$

The logistic regression model employed in this analysis assumes that the relationship between the predictors and the binary outcome (donor classification) is approximately linear on the log-odds scale. This assumption implies that the effect of each predictor on the log-odds of being a donor is consistent across different levels of the predictor variables. Additionally, logistic regression assumes the absence of multicollinearity among predictors, ensuring that each predictor contributes uniquely to the model's predictive power.

Multiple Linear Regression 1:

The assumptions for multiple linear regression include linearity, where the relationship between predictors and the response variable is assumed to be linear. Additionally, normality of residuals, absence of perfect multicollinearity, and no autocorrelation among residuals are also assumed for the model to be valid. Violations of these assumptions may affect the reliability and accuracy of the regression analysis.

$$\begin{aligned}\text{Donor\_AMT} =& \beta_0 + \beta_1 \text{region\_2} + \beta_2 \text{home\_owner} + \beta_3 \text{number\_children} - \\ & + \beta_4 \text{wealth\_rating} + \beta_5 \text{income} + \beta_6 \text{avg\_family\_income} \\ & + \beta_7 \#\text{ of promotions} + \beta_8 \text{number of months between first and second gift}\end{aligned} \tag{2}$$

Results

Based on the logistic regression model's classification table Table 3, which correctly predicted 387 instances of sending mail (1) and 1620 instances of not sending mail (0), we conclude that the decision to send mail to the 387 individuals with the highest posterior probabilities aligns with maximizing potential response. The model's performance is further highlighted by the optimal number of mailings (1378), resulting in a maximum profit of $11,584.5. In contrast, our linear regression model yielded a mean squared error (MSE) of 3.75, with a standard error of 0.262, indicating some variability in prediction accuracy. However, the linear regression model's predictive power is not as strong as the logistic regression's classification ability, suggesting that the latter may be more effective in maximizing response rates and profitability.

5 Conclusion

In conclusion, our analysis demonstrates the significance of data-driven methodologies in refining fundraising strategies for charitable initiatives. By employing generalized linear models (GLMs), we've discerned influential factors governing donor behavior and donation magnitudes. Notably, our logistic regression model showcased superior predictive accuracy, aiding in the identification of potential donors and optimizing response rates. Through this model, we determined that sending mail to the top 387 individuals with the highest posterior probabilities maximizes potential response, resulting in an optimal number of 1378 mailings and a net profit of $11,584.5. Additionally, our linear regression model, although less robust in predictive power, provides valuable insights into donation amount prediction. Integrating these findings into future campaigns promises to enhance cost-effectiveness and amplify the impact of charitable endeavors.

6 Appendix

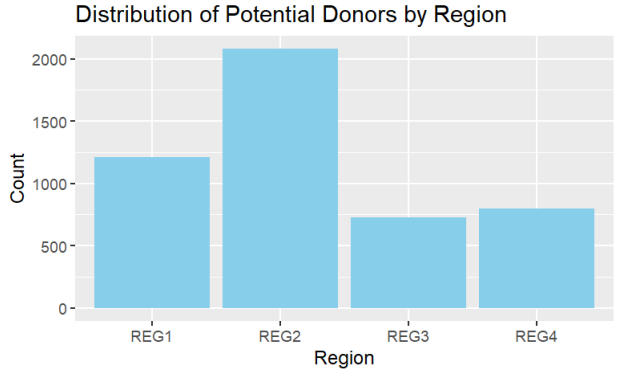

Figure 1: Count of possible donors per region

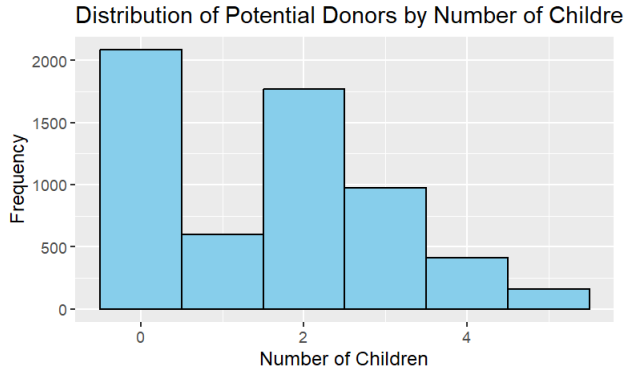

Figure 2: Histogram of children

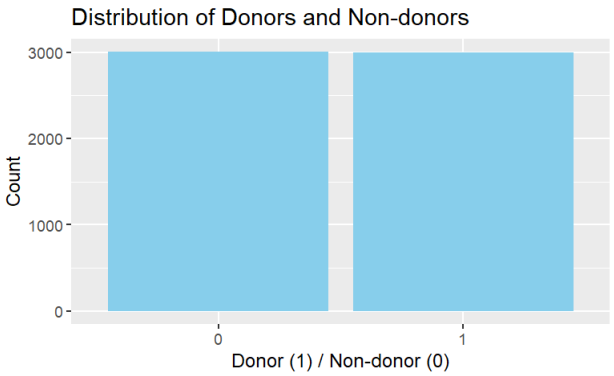

Figure 3: Counts of Donors (1) vs Not donors (0)

Figure 4: Counts of Donors and not donors separated by gender

Variable	**Correlation (donr)**
home	0.2916
reg2	0.2534
damt	0.0267
chld	-0.532
PLOW	-0.131
wrat	0.238
npro	0.139
income	0.1394

Table 1: Correlation with being classified as a Donor

Variable	Correlation (damt)
home	0.2895
reg2	0.2137
donr	0.9817
chld	-0.55
npro	0.146
incm	0.1452
wrat	0.231

Table 2: Correlation with how much one Donates

	Predicted 0	**Predicted 1**
Actual 0	630	10
Actual 1	389	989
Total	1019	999

Table 3: Classification Table for Logistic Regression Model