Project: What Gives?

Members: Vanessa Ma, Shyamsunder Sriram

Executive Summary

Paralyzed Veterans of America (PVA) will benefit from utilizing algorithmic targeting methods to solicit donations through direct mailing campaigns. Deploying our model will increase the profit of their upcoming mailing campaign by $2,139 to $14,468, a 17.36% increase.

Objective: Can PVA increase philanthropic returns by selectively mailing?

Thus, a target model must therefore be able to accurately predict:

1. Who, out of the existing database, will respond in an upcoming mailing?
2. Among those who are predicted to respond, how much will they give?

Theory & Model

We theorize that the above two questions are actually two different utility decisions for an individual donor, and the full utility of philanthropy is an addition of the two components.

$$U_i(X_i) = U_i{}^1(X_i{}^1) + U_i{}^2(X_i{}^2)\, s.t.\, X_i{}^1,\, X_i{}^2 \in X_i;\; U^1 \neq U^2$$

The first term is the utility of donating or not, a similar choice to general consumption, as it is functionally not different from committing to a "method" of spending money - he decides to donate if committing to philanthropy gives him more utility than not doing so. This is a binary decision, which we hypothesize is largely driven by a consumer's connection to the organization. So the corresponding utility function for donor i would be:

$$U_i{}^1(X_i{}^1) = \alpha_i{}^1 + \beta_i{}^1 X_i{}^1 + \varepsilon$$

- $X_i{}^1$: characteristic set that predicts a donor's utility with respect to the binary choice of engaging in philanthropy.

We needed a binary classification model, and chose the logistic regression for its simplicity.

Now let us consider the second term, the utility of the donation amount itself. Unlike the first decision, the resulting variable from this action is continuous, and dissimilar to a consumption

decision - the price is not fixed, and is freely assigned by the targeted donor. The donor decides to donate this exact amount if neither one more or less dollar would gain him more utility if spent on a myriad of other charities. Thus, this is a question of regression. We further hypothesize that this decision is largely driven by a consumer's connection to the organization, such that the greater the personal connection, the more utility he would gain by spending this dollar on PVA compared to any other organization. The corresponding utility function for donor i:

$$U_i^2(X_i^2) = \alpha_i^2 + \beta_i^2 D(X_i^2) + \varepsilon$$

- X_i^2 : characteristic set that predicts a donor's utility with respect to the continuous choice of how much to give
- $D(...)$: function that outputs giving amount based on individual characteristics

A model that answers this question must therefore be a regression model. We chose a generalized linear model for its simplicity and flexibility in taking in large numbers of variables.

Data

Exploratory data analysis was conducted with the goal of crafting a target measure of success. Our data is not a randomized experiment, shown by propensity scores skewed to 1, nor are the small non-treated set comparable to the treated population for any past mailing, shown by covariate imbalance. Thus, we cannot use the typical baseline, which is spend without marketing. Instead, we picked our baseline to be a blanket mailing strategy, calculated as the total of all donations predicted, net of blanket mailing costs (i.e. total targeted donors * $0.68, the flat rate of mailing).

Our final profit model was thus defined as follows:

$$\pi = \sum_{x,y} f(x,y) = f_2(y|x) * f_1(x) - z \; s.t. \; x \in X, \; y \in Y$$

- X be a random variable that represents whether a prospective donor donates

Project: What Gives?
Members: Vanessa Ma, Shyamsunder Sriram

- Y be a random variable that represents the value of the donation given
- Z be the cost of mailing
- $f_1(X)$ is our logistic regression model
- $f_2(Y \mid X = 1)$ from our linear regression model

Methods

Due to the sheer size of the dataset, we created a “quick and dirty” stepwise method to isolate key predictors for the logit and linear regression models: we initially ran the full model in the regression and selected predictors had p-values < 0.08, and ran a regression again among those selected predictors. We repeated this process 5 times, and chose the threshold with the 5% alpha value in mind in order to prevent an overly conservative bound. Since our goal is prediction, we don’t mind having a few substandard predictors with a model with mostly significant predictors, but we want to capture all of our prime predictors.

Implementation

Logistic Regression to Predict Responses

After implementing ”Quick Stepwise” we got a result of 31 predictors that incorporated select data from demographics, census, giving history, and past mail order responses. We calculated the predicted logit values on the validation dataset and organized them by decile. We organized the logit values by decile and captured the amount of donors captured in our validation dataset. We found that the % Donors captured generally decreases as we lower our threshold, which indicates stability in our model.

Linear Regression to Predict Donations

15 predictors were gleaned from the “Quick Stepwise” method for the linear regression, incorporating data from past giving history, some census data. To prepare the data to fit with linearity constraints, we removed predictors that were highly correlated and had high

collinearity, ensured that there was limited correlation between residuals and predictors (checking residual plots) and removed outlier points by checking studentized residuals. The resulting model had an R-squared value of 0.8166, and an adjusted R-squared value of 0.8159.

Final Profit Predictions

Our targeting scheme will target consumers with a probability of donating above a certain threshold probability from the logit model. Each decile corresponds to a logit threshold. Our profit formula thus came down to:

$$\textbf{Profit} = T * (LP * D - C)$$

- LP: Logit Probabilities
- D: Predicted Donations
- T: targeting indicator (if LP > 0 then 1 else 0)
- C: cost of mailing promotion

On the validation set, we calculated the maximum profits by targeting the top 80% of donors, predicting $2749.35 profits compared to actual profits of $2664.74.

Using the same methodology on the test set, we were able to achieve the following results[1]:

	KDD Testing Set* Actual Results (provided by organization)	Our Predicted Results
Net Profits	$14,712	**$14,468**
Baseline Profits if entire population is targeted	$10,560	**$12,329**
Increase in Net Donation	$4,152	**$2,139**
% Increase	39.32%	**$17.35%**

While our model underestimated the baseline profits, it had accurate results for the testing dataset, as well as our validation dataset. Hence, we can conclude that this model is a reliable predictor for future donations.

[1] * https://www.kdnuggets.com/meetings-past/kdd98/gain-kddcup98-release.html (KDD Actual figures taken from the winner's site.)