## Direct Mail Fundraising

**Background:**

A national veterans organization[1] wishes to develop a data-mining model to improve the cost-effectiveness of their direct marketing campaign. The organization, with its in-house database of over 13 million donors, is one of the largest direct mail fundraisers in the United States. In a recent mailing to a small portion of the general list, the overall response rate was 5.1%. Out of those who responded (donated), the average donation is $13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $0.68 to produce and send. Using these facts, we take a sample of this data set to develop a classification model that can effectively capture donors so that the expected net profit is maximized. This model can then be applied to future data. Weighted sampling is used, over-representing the responders so that the sample has equal numbers of donors and non-donors.

**Data:**

The file "DonorData.xls" contains 3120 data points with 50% donors (TARGET_B=1) and 50% non-donors (TARGET_B=0). The amount of donation (TARGET_D) is also included but is not used in this case. The descriptions for the 25 attributes (including two target variables) are listed as follows:

| | |
|---|---|
| ZIP | Zipcode group (zipcodes were grouped into 5 groups; only 4 are needed for analysis since if a potential donor falls into none of the four he or she must be in the other group. Inclusion of all five variables would be redundant and cause some modeling techniques to fail. A "1" indicates the potential donor belongs to this zip group.)<br>00000-19999 => 1 (omitted for above reason)<br>20000-39999 => zipconvert_2<br>40000-59999 => zipconvert_3<br>60000-79999 => zipconvert_4<br>80000-99999 => zipconvert_5 |
| HOMEOWNER | 1 = homeowner, 0 = not a homeowner |
| NUMCHLD | Number of children |
| INCOME | Household income |
| GENDER | Gender<br>0 = Male<br>1 = Female |
| WEALTH | Wealth Rating<br>Wealth rating uses median family income and population statistics from each area to index relative wealth within each state<br>The segments are denoted 0-9, with 9 being the highest wealth group and zero being the lowest. Each rating has a different meaning within each state. |
| HV | Average Home Value in potential donor's neighborhood in $ hundreds |
| ICmed | Median Family Income in potential donor's neighborhood in $ hundreds |
| ICavg | Average Family Income in potential donor's neighborhood in hundreds |
| IC15 | Percent earning less than 15K in potential donor's neighborhood |
| NUMPROM | Lifetime number of promotions received to date |

[1] The name of the organization cannot be revealed for proprietary reasons.

```
RAMNTALL          Dollar amount of lifetime gifts to date
MAXRAMNT          Dollar amount of largest gift to date
LASTGIFT          Dollar amount of most recent gift
TOTALMONTHS       Number of months from last donation to July 1998 (the last time
                  the case was updated)
TIMELAG           Number of months between first and second  gift
AVGGIFT           Average dollar amount of gifts to date
TARGET_B          Target Variable: Binary Indicator for Response
                  1 = Donor
                  0 = Non-donor
TARGET_D          Target Variable: Donation Amount (in $).  We will NOT
                  be using this variable for this case.
```

**Step 1: Partitioning**

Partition the dataset into 60% training and 40% validation (set the seed to 12345).

**Step 2: Model Building**

**(a) Selecting classification tool and parameters**

Run the following classification tools on the data:

- Logistic Regression
- Classification Trees
- Neural Networks

Be sure to test different parameter values for each method. You may also want to run each method on a subset of the variables. (Note: due to a glitch, the "subset selection" routine of logistic regression may not work for this particular data set; to choose a subset of variables in logistic regression for this problem, we suggest you use the p-values of the predictor variables.) Be sure NOT to include "TARGET_D" in your analysis.

**(b) Classification under asymmetric response and cost**

What is the reasoning behind using weighted sampling to produce training and validation sets with equal numbers of donors and non-donors? Why not use a simple random sample from the original dataset? (Hint: given the actual response rate of 5.1%, how do you think the classification models will behave under simple sampling)? In this case, is classification accuracy a good performance metric for our purposes of maximizing net profit? If not, how would you determine the best model? Please explain your reasoning.

**(c) Calculate Net Profit**

For each method, calculate the lift of net profit for both the training and validation set based on the actual response rate (5.1%). Again, the expected donation, given that they are donors, is $13.00, and the total cost of each mailing is $0.68. (Hint: to calculate estimated net profit, we will need to "undo" the effects of the weighted sampling, and calculate the net profit that would reflect the actual response distribution of 5.1% donors and 94.9% non-donors.)

**(d) Draw Lift Curves**

Draw each model's net profit lift curve for the validation set onto a single graph. Are there any models that dominate?

**(e) Best Model**

From your answer in part 2b, what do you think is the "best" model?