IDS 572 – Assignment 2 Target Marketing - Fundraising
Due: Feb 21st, 2016

Background

A national veteran's organization 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. According to their recent mailing records, the overall response rate is 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 dataset to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling is used, under-representing the non-responders so that the sample has a more balanced numbers of donors and non-donors.

Data
The file pvaBalanced35Trg.csv contains 9999 data points. The sample has been balanced to carry 35% donors i.e.the data has 35% donors (TARGET−B = 1) and 65% non-donors (TARGET−B = 0). The amount of donation (TARGET−D) is also included but is not used in this case. The file contains all 480 attributes.

Assignment

In this assignment, we need to clean the data, conduct an exploratory analysis on which variables may be useful to predict donors, and then build a predictive model.

1. The dataset has many variables – some (most?) of them may not be useful for our purpose. Your first task is to clean and explore the data, determine missing values and how you might handle these, which variables you think need not be considered, which should be transformed, etc. This is a major task – and can take time, much more than the modeling step that comes next. You will find below a list of subset of variables that someone found useful. Which variables will you consider for modeling (and why)?
Which attributes will you omit from the analyses and why.
How do you clean the data, handle missing values? What new attributes/values do you derive?.
How do you approach data reduction? What methods for data reduction do you try?

Data cleaning - certain variables have 'empty' values in many rows. Some of these may be actual missing values, while the empty values may carry information (e.g. for a variable like collegeEducation, empty values may indicate no-college-education which can be coded as a specific value). Some variables carry separate information in different bytes.....
Outline the data cleaning steps that you perform (and why)

Data exploration: Import the data, and examine the different variables – distribution of values, mean and std deviation, range of values. What do you observe?
What variable transformations do you make (and why)?
Perform Principal Components Analysis (PCA) – which variables do you include for PCA (give your reason).
Do decision trees help determine which variables to include in a predictive model for donors? How?

2. Modeling

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

[A specified seed ensures that we obtain the same random partitioning every time we run it. With no specified seed, the system clock is typically used to set the seed, and a different partitioning can result in different runs].

Consider the following classification techniques on the data:

- decision Trees (you can use J48, or any other suitable type of decision tree)
- logistic Regression
- naïve-Bayes

Be sure to test different parameter values for each method, as you see suitable. What parameter values do you try for the different techniques, and what do you find to work best?

Run each method on a chosen subset of the variables - how do you select this subset?

(Be sure NOT to include "TARGET−D" in your analysis.)

Provide a comparative evaluation of performance of your best models from each technique.

Does variable selection/PCA make a difference for the different models?

3. Classification under asymmetric response and cost: What is the reasoning behind using weighted sampling to produce a training set 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? Explain your reasoning.

Table 13.7: Description of Variables for the Fundraising Dataset

HOMEOWNER 1 = homeowner, 0 = not a homeowner
NUMCHLD Number of children
INCOME Household income
GENDER 0 = Male, 1 = Female
WEALTH Wealth Rating
Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. 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 (IC2)
ICavg Average Family Income in potential donor's neighborhood in hundreds (IC4)
IC15 Percent earning less than 15K in potential donor's neighborhood
NUMPROM Lifetime number of promotions received to date
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.