IDS 572 – Assignment 3   Target Marketing – Fundraising (Part 2)
Due: Oct 17th, 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

This is the continuation of Assignment 2, where you conducted an exploratory data analysis, data cleaning, data reduction,…, and developed some classification models to predict donors (TARGET_B). In this assignment, we also examine the performance of nearest neighbors and support vector machines for classifying donors. Performance should be evaluated based on costs and benefits as given above; we want a model that maximizes profit. We will also develop a model to predict donation amount (TARGET_D) directly, and examine how to combine the response and donation amount models to identify the most profitable individuals to target.

1. Modeling
Partitioning - Partition the dataset into 60% training and 40% validation (set the seed to 12345).
In the last assignment, you developed decision tree, logistic regression, naïve Bayes, random forest and boosted tree models. Now, develop support vector machine models for classification. Examine different parameter values, as you see suitable. Report on what you experimented with and what worked best.

How do you select the subset of variables to include in the model? What methods do you use to select variables that you feel should be included in the model(s)? Does variable selection make a difference?

Provide a comparative evaluation of performance of your best models from all techniques (including those from part 1, ie. assignment 2)

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

2. Our overall goal is to identify which individuals to target for maximum donations (profit). We will try two approaches for this:

(i) using the response model, together with average donation and mailing costs information, to identify the most profitable individuals to target (Q 2.1 below)

(ii) develop a second model on TARGET_D, and combine this with the response model to identify the most profitable individuals to target (Q 2.2 below)

2.1 (a) What is the 'best' model for each method in Question 1 for maximizing revenue? Calculate the net profit for both the training and validation set based on the actual response rate (5.1%). We can calculate the net profit from given information - the expected donation, given that they are donors, is $13.00, and the total cost of each mailing is $0.68.
Note: to calculate estimated net profit (on data with the 'natural' response rate of 5.1%), we will need to "undo" the effects of the weighted sampling, and calculate the net profit that reflects the actual response distribution of 5.1% donors and 94.9% non-donors.)

(b) Summarize the performance of the 'best' model from each method, in terms of net profit from predicting donors in the validation dataset; at what cutoff is the best performance obtained?

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

Best Model: From your answers above, what do you think will be the "best" model to implement? (What criteria do you use to determine 'best'?)

2.2. (a) We will next develop a model for the donated amount (TARGET_D). Note that TARGET_D has values only for those individuals who donors (that is, TARGET_D values are defined only for cases where TARGET_B is 1).
What data will you use to develop a model for TARGET_D? (Non-donors, obviously, do not have any donation amount -- should you consider these as $0.0 donation, or impute missing values here? Should non-donors be included for developing the model to predict donation amount? Also, should cases with rare very large donation amounts be excluded? [Hah ! – leading questions☺]

Develop a model for TARGET_D. What modeling method do you use (report on <u>any one</u>).
Which variables do you use? What variable selection methods do you use? Report on performance.

(b) How can you use the results from the response model together with results from the donation amount model to identify targets?
(Hint:The response model estimates the probability of response p(y=1|x). The donation amount model estimates the conditional donation amount, E[amt | x, y=1]. The product of these gives …..?)

How do you identify individuals to target, when combining information from both models?
[Sorted values of predicted donation amount? What threshold would you use? Or maybe you prefer to target all individuals with predicted amounts greater than $0.68 (why?), and/or…..]

5. <u>Testing</u> – chose one model, either the one from 2.1 or 2.2 above, based on performance on the test data.

The file FutureFundraising.xls contains the attributes for future mailing candidates. Using your "best" model from Step 2 which of these candidates do you predict as donors and non-donors? List them in descending order of probability of being a donor/prediction of donation amount. What cutoff do you use? Submit this file (xls format), with your best model's predictions (prob of being a donor).

[The FutureFundraising.xls data file does not contain values of the target variable – so you cannot really see how your model performs on this data. In evaluating your assignment, the instructor will determine how your 'best' model performs on this data. Note that part of each team's evaluation will be based on how well you model performs relative to other models submitted by other teams.]