ADS-505 Applied Data Science for Business

Module 4 Assignment 4.1: Use Case - Direct-Mail Fundraising (Using Python) (80 points)

Fundraising.csv and Futurefundraising.csv are the datasets for this case study (linked in the assignment prompt in Blackboard).

For this assignment, you will use a Jupyter notebook to answer the questions in the "Assignment" section of this document. Be sure to include the number and question that you are answering in your notebook.

BACKGROUND

Note: Be sure to read the information about oversampling and adjustment in Chapter 5 before starting to work on this case.
A national veterans' organization wishes to develop a predictive 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 equal numbers of donors and non-donors.

DATA

The file *Fundraising.csv* contains 3120 records 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 22 variables (including two target variables) are listed in Table 21.9.

ASSIGNMENT

1. ***Data preparation**:* Load the data and prepare it for predictive analysis.

   1.1 Partition the dataset into 60% training and 40% validation (set the seed to 12345). **(4 Points)**

2. ***Model Building***: Follow the following steps to build, evaluate, and choose a model.

   2.1. *Select classification tool and parameters*: Run at least two classification models of your choosing. Be sure NOT to use TARGET_D in your analysis. Describe the two models that you chose, with sufficient detail (method, parameters, variables, etc.) so that it can be replicated. **(10 Points)**

   2.2. *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? **(10 Points)**

   2.3. *Calculate net profit*: For each method, calculate the cumulative gains of net profit for both the training and validation sets 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. To do this, divide each row's net profit by the oversampling weights applicable to the actual status of that row. The oversampling weight for actual donors is 50%/5.1% = 9.8. The oversampling weight for actual non-donors is 50%/94.9% = 0.53.) **(10 Points)**

Table 21.9 Description of Variables for the Fundraising Dataset

Variable	Description
ZIP	Zip code group (zip codes were grouped into five groups;
	1 = the potential donor belongs to this zip group.)
	00000–19999 ⇒ zipconvert_1
	20000–39999 ⇒ zipconvert_2
	40000–59999 ⇒ zipconvert_3
	60000–79999 ⇒ zipconvert_4
	80000–99999 ⇒ zipconvert_5
HOMEOWNER	1 = homeowner, 0 = not a homeowner
NUMCHLD	Number of children

INCOME	Household income
GENDER	0 = male, 1 = female
WEALTH	Wealth rating uses median family income and population statistics
	from each area to index relative wealth within each state
	The segments are denoted 0 to 9, with 9 being the highest-wealth
	group and zero the lowest. Each rating has a different
	meaning within each state.
HV	Average home value in potential donor's neighborhood in hundreds of dollars
ICmed	Median family income in potential donor's neighborhood in hundreds of dollars
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
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*	Outcome variable: binary indicator for response
	1 = donor, 0 = non-donor
TARGET_*D*	Outcome variable: donation amount (in dollars). We will NOT be using this variable for this case.

2.4. *Draw cumulative gains curves*: Draw the different models' net profit cumulative gains curves for the validation set in a single plot (net profit on the *y*-axis, proportion of list or number mailed on the *x*-axis). Is there a model that dominates? **(10 Points)**

2.5. *Select the best model*: From your answer in (4), what do you think is the "best" model? **(12 Points)**

3. ***Testing***: The file *FutureFundraising.csv* contains the attributes for future mailing candidates.

   3.1. Using your "best" model from Step 2 (number 5), which of these candidates do you predict as donors and non-donors? List them in descending order of the probability of being a donor. Starting at the top of this sorted list, roughly how far down would you go in a mailing campaign? **(10 Points)**

   3.2. Briefly explain, in two to three paragraphs, the business objective, the data mining models used, why they were used, the model results, and your recommendations to your non-technical stakeholder team. **(14 Points)**