Case Study: A Fund Raising Net Return Prediction Model

Background and objectives

"Healthcare for All", is a not-for-profit organization that provides financial help to people who are not able to afford healthcare. "Healthcare for All" raises funds through donations from all across the country. They have an in-house database of over 13 million donors. "Healthcare for All" has been consistently raising money through various campaigns to request people to contribute to their cause. They reach out to the donors through various channels including personal mails, emails, fundraising events, reaching out to other businesses and corporations, and other philanthropists. One of their most efficient channels for a long period of time has been direct mails. But for the last couple of years, they have seen a decline in the donations through this medium.

You are working for "Healthcare for All" as an analyst. You will be analyzing the results of one of their recent direct mail fundraising appeals. This mailing was sent to a total of 3.5 million "Healthcare for All" donors who were on the database as of June 1997. Everyone included in this mailing had made at least one prior donation.

The mailing included a gift (or "premium") of personalized name & address labels plus an assortment of 10 note cards and envelopes. All of the donors who received this mailing were acquired by "Healthcare for All" through similar premium-oriented appeals such as this.

One group that is of particular interest to the organization is "Lapsed" donors. These are individuals who made their last donation 13 to 24 months ago. They represent an important group to the organization, since the longer someone goes without donating, the less likely they will be to give again. Therefore, the recapture of these former donors is a critical aspect of their fund raising efforts.

However, it was found that there is often an inverse correlation between likelihood to respond and the dollar amount of the gift, so a straight response model (a classification or discrimination task) will most likely net only very low dollar donors. High dollar donors will fall into the lower deciles, which would most likely be suppressed from future mailings. The lost revenue of these suppressed donors would then offset any gains due to the increased response rate of the low dollar donors.

Therefore, to improve the cost-effectiveness of future direct marketing efforts, "Healthcare for All" wishes to develop a model that will help them maximize the net revenue (a regression or estimation task) generated from future renewal mailings to Lapsed donors.

Population

The population for this analysis will be Lapsed donors who received the June 1997 renewal mailing. Therefore, the analysis data set contains a subset of the total universe that received the mailing. The analysis file includes all 191,779 Lapsed donors who received the mailing, with responders to the mailing marked with a flag in the TARGET_B field. The total dollar amount of each responder's gift is in the TARGET_D field.

The overall response rate for this direct mail promotion is 5.1%. The distribution of the target fields in the learning and validation files is as follows:

Learning Data Set Target Variable: Binary Indicator of Response to Mailing

TARGET_B	Frequency	Percent
0	90569	94.9
1	4843	5.1

Target Variable: Donation Amount (in \$) to Mailing

Variable	N	Mean	Minimum	Maximum
TARGET_D	95412	0.7930732	0	200.0000000

Validation Data Set Target Variable: Binary Indicator of Response to Mailing

TARGET_B	Frequency	Percent
0	91494	94.9
1	4873	5.1

Target Variable: Donation Amount (in \$) to Mailing

Variable	N	Mean	Minimum	Maximum
TARGET_D	96367	0.7895819	0	500.0000000

Cost Matrix

The package cost (including the mail cost) is \$0.68 per piece mailed.

Analysis time frame and reference date

The mailing was sent out in June 2018. All information included in the file (excluding the giving history date fields) is reflective of behavior before 6/97. This date may be used as the reference date in generating the "number of months since" or "time since" or "elapsed time" variables. You can also find the reference date information in the filed `ADATE_2`. This filed contains the dates the promotion was mailed.

Data Sources and Order & Type of the variables in the data set

The dataset includes:

- 24 months of detailed promotion and giving history (covering the period 12 to 36 months before the mailing)
- A summary of the promotions sent to the donors over the most recent 12 months before the mailing (by definition, none of these donors responded to any of these promotions)
- Summary variables reflecting each donor's lifetime giving history (e.g., total # of donations before mailing, total \$ amount of the donations, etc.)
- Overlay demographics, including a mix of household and area level data
- All other available data from the database (e.g., date of first gift, state, origin source, etc.)

SOME CHALLENGES WITH THE DATASET

The dataset provided in this case study is not an easy dataset to manage. There are a few reasons for it:

- First, there is a large number of features provided. The data set has over 450 features. Hence, selecting the right features for the model is very critical and at the same time it is not easy as the same traditional ways of removing features is not effective given the large number of features. Apart from feature selection, feature extraction (creating your own features using the existing features) is also not easy in this case.
- Sparsity of the dataset is another issue. There are a lot of features with a large number of null values.
- Data imbalance - For developing a classification, there is a huge imbalance in the training dataset with only approximately 5000 values for one category as compared to over 95,000 instances for the other category.