ESSEC
BUSINESS SCHOOL

MARKETING ANALYTICS

Assignment #2

ARNAUD DE BRUYN

Overview

Assignment #2 is a predictive modeling exercise.

Your goal is to predict who is likely to make a donation to the charity for the fundraising campaign "C189" (%), and how much money they are likely to give if they do (€). By combining these two predictions (% × €), you will obtain an expected revenue from each individual.

Every solicitation costs 2.00 € (a fake, unrealistic figure used for the purpose of this exercise).

If the expected revenue you have predicted exceeds that figure of 2 €, you will recommend the charity to solicit that individual (solicit = 1), since the expected profit is positive. If it is below 2 €, you will recommend the charity *not* to solicit that individual (solicit = 0), since on average you expect a loss.

For the purpose of this exercise, we will assume that no individual is going to make a donation on that campaign if he is not directly solicited by the charity. We will also ignore all donations made under automatic deductions following that campaign (they will not be included in the data).

Your objective is to maximize the financial performance of that campaign for the charity.

Logistics

Deadline

The deadline is on Sunday at midnight. Please check the slide deck of session 1 for the exact date.

Data set

The data set used for this exercise is the "ma_charity_full" database (refer to the database course document for details). The file is quite voluminous (85 Mb zipped, 541 Mb unzipped, 617 Mb once loaded as an MySQL database). Once you have downloaded and unzipped the file, please check the document called "database description," which contains further instructions on how to load the database on your MySQL database server.

Submission

You have to prepare a text file in a very specific format (see instructions below), and post this file on the following URL, along with a unique code that will be submitted to you by email:

http://www.debruyn.info/essec/ma/

Do NOT submit solutions before you have been instructed to do so by the instructor, even if the website appears to be available!

You can submit up to 3 files with different recommendations, as long as it is before the deadline.

As soon as you submit a file, your financial performance will be computed and communicated to you.

Financial performance will be computed based on actual behavior observed in the database for that campaign, and known by the professor only.

Grade

Your financial performance will be compared to the financial performance of your peers. If you have submitted multiple solutions, only the one with best performance will be retained (your best performance might not necessarily correspond to your last submission).

The student with the highest financial performance will receive a grade of 20/20. The student with the worst financial performance will receive a grade of 8/20. All the other students will be ranked and receive a grade commensurate to their rank. For instance, students at the 20th top percentile will receive 17.6/20. By design, the average grade for this assignment will be 14/20.

Students who do not submit anything will receive a 0/20 for the assignment, and will be counted in the pool. Consequently, the lowest grade might start above 8/20.

The professor will also compete under the same conditions as the students (blind recommendations, 3 essays maximum, etc.). Any student who achieves a financial performance superior to the professor's will receive a +1 bonus point for the entire course. This bonus point will be added to the overall grade of the course, not simply to the grade of this assignment. It is therefore technically possible to obtain 21/20 for this course (although the total grade is capped at 20 in the system).

Honor code

The assignment is individual. You are expected to work on it by yourself. Suspected cheating will grant you an automatic 0/20 for the assignment.

The code that will be sent to you to submit your work is also individual. You should not share it.

R code

Between the deadline and the next session, you are requested to send your code by email to debruyn@essec.edu.

Please note that the assignment has to be done in R. Students who fail to submit their R scripts, or who fail to submit any code at all, will be heavily penalized, and will automatically fail the assignment. While I use Python or SPSS regularly, and I am sympathetic with the relative strengths and weaknesses of numerous programming languages, the official programming language of this course is R. No exception.

Technical details

Data set

On the last day of data available in the database, the charity has solicited 123 672 donors for their June campaign.

Out of these 123 672 solicited donors, about ~12 700 have made a subsequent donation. Which contacts have decided to make a donation, for how much, and who did not, has been observed and is known by the professor.

These 123 672 solicited donors have been divided into two batches of approximately equal sizes.

For the first batch (N = 61 928), called the “calibration” data, you have complete information about their responses to the fundraising campaign. This batch will be used for calibration.

For the second batch (N = 61 744), called the “prediction” data, you only know that they have been solicited, but their actual responses have not been communicated to you, and have been excluded from the data you have received. This batch will be used for performance evaluation.

All that information is contained in the “assignment2” table:

- contact_id The list of individuals who have been solicited by the charity for this specific fundraising campaign.
- calibration 1 if part of the calibration data, 0 if part of the prediction data.
- donation 1 if the donor has made a donation, 0 if the donor has not, NULL if the information is unknown to you. By design, the value is set to NULL if calibration = 0.
- amount The actual donation amount observed, in EUR. The value is NULL if the donor has not made any donation (donation = 0) *or* if the response has not been communicated to you for this individual (calibration = 0).
- act_date The date at which the donation has been made. The value is NULL if the donor has not made any donation (donation = 0) *or* if the response has not been communicated to you for this individual (calibration = 0).

Process

Following the procedures we have explored and explained in the course, you need to:

1. Calibrate a discrete model (%) to predict the likelihood of donation (on individuals where calibration = 1)
2. Calibrate a continuous model (€) to predict the most likely donation amount in case of donation (on the subset of individuals where donation = 1)
3. Apply both models to the prediction data (i.e., individuals where calibration = 0), and multiply these predictions (% and €) to obtain expected (predicted) revenue if solicited.
4. If expected revenue is superior to 2.00 €, solicit (=1); otherwise do not (=0).

File format

The file you will submit needs to follow these specifications precisely. Failure to do so will prevent you from submitting your recommendations:

- Text file
- TAB separator
- No header
- Exactly 61 744 lines (the number of individuals with calibration = 0)
- First column contains the contact_id
- Second column contains a 1 (solicit) or a 0 (do not solicit)
- Make sure to sort your recommendations by contact_id

For instance:

```
153	1
185	0
283	1
319	0
329	0
371	1
…
```

Usual errors

Errors to avoid:

- Use “space” as separator rather than a TAB
- Include two TABs instead of one (this creates a second column, all empty, which is interpreted as “0” everywhere)
- Submit a file that does not include exactly 61 744 lines
- Submit a file where ids are not listed in increasing order