MATH 80619 Advanced statistical learning

Assignment #6 – uplift modeling and prediction intervals

Specific instructions

- In order to ensure reproducibility of your results, your assignment must be written using R Markdown or R Sweave. Your R code, output and comments should be included in the main document.
- You can submit the html or pdf output file on Zone Cours in the section "Remise de Travaux".
- Deadline is Monday, April 19th, before 5 pm.

Question #1

In this question, you will analyze one of the few public datasets available for uplift modeling in the package `Information`. The data is from a historical marketing campaign. It contains 68 predictive variables including a treatment indicator and a purchase indicator. The definition of all these variables is hard to find but the most relevant ones for your analysis are:

- TREATMENT: equals 1 if the person received the marketing offer, and 0 if the person was in the control group
- PURCHASE: equals 1 if the person accepted the offer, and 0 otherwise
- UNIQUE_ID: unique identifier
- AGE: age of the person
- D_REGION_X: 1 if the person lives in region X, 0 otherwise (3 regions: A, B, C)

Other variables are from credit bureau data (e.g., N_OPEN_REV_ACTS = number of open revolving accounts).
The train and test datasets are in the objects `uplift.train` and `uplift.test` in the file `uplift.RData`.

1) Use the two-model approach using a logistic regression and all variables in the model to obtain the lift for all subjects in the test set. Print the lift for the top 10 subjects.

2) Repeat the same question by using a forest with 100 trees. Compare the ranks of the lift for both models.

3) Repeat the same question by using a class transformation using i) a logistic model and ii) a forest with 100 trees. You can assume a propensity score of 0.5.

4) Compare all models using the Qini coefficient.

5) Comments on the results and give your thought about this analysis.

Question #2

Create 2 questions of the type “multiple choice questions” on the material seen in the chapter on prediction intervals. **Please, don’t use questions based on simple definitions.**

- Your questions should have 4 possible answers, 1 true answer and 3 false answers.
- The level of the question should not be too easy (i.e. wrong answers should not be obvious).
- Solutions - with explanations for the right and wrong answers- need also to be included.

Your mark for this question will be based on the correctness of your question/answers, and on the depth of your question and answers (especially the wrong ones).