COURSE DETAILS

Instructor Name: Dr. Nathan Bastian (Prof. B)
Course Name: Big Data Econometrics
Course Number: ADEC 7430

COURSE PROJECT DESCRIPTION

A charitable organization wishes to develop a machine learning model to improve the cost-effectiveness of their direct marketing campaigns to previous donors. According to their recent mailing records, the typical overall response rate is 10%. Out of those who respond (donate) to the mailing, the average donation is $14.50. Each mailing costs $2.00 to produce and send; the mailing includes a gift of personalized address labels and assortment of cards and envelopes. It is not cost-effective to mail everyone because the expected profit from each mailing is 14.50 x 0.10 – 2 = -$0.55. We would like to develop a classification model using data from the most recent campaign that can effectively captures likely donors so that the expected net profit is maximized. The entire dataset consists of 3984 training observations, 2018 validation observations, and 2007 test observations. Weighted sampling has been used, over-representing the responders so that the training and validation samples have approximately equal numbers of donors and non-donors. The response rate in the test sample has the more typical 10% response rate. We would also like to build a prediction model to predict expected gift amounts from donors – the data for this will consist of the records for donors only. The data are available in the file "charity.csv" (available in Canvas):

- ID number [Do NOT use this as a predictor variable in any models]
- REG1, REG2, REG3, REG4: Region (There are five geographic regions; only four are needed for analysis since if a potential donor falls into none of the four he or she must be in the other region. Inclusion of all five indicator variables would be redundant and cause some modeling techniques to fail. A "1" indicates the potential donor belongs to this region.)
- HOME: (1 = homeowner, 0 = not a homeowner)
- CHLD: Number of children
- HINC: Household income (7 categories)
- GENF: Gender (0 = Male, 1 = Female)
- WRAT: Wealth Rating (Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and 0 being the lowest.)
- AVHV: Average Home Value in potential donor's neighborhood in $ thousands
- INCM: Median Family Income in potential donor's neighborhood in $ thousands
- INCA: Average Family Income in potential donor's neighborhood in $ thousands
- PLOW: Percent categorized as "low income" in potential donor's neighborhood
- NPRO: Lifetime number of promotions received to date
- TGIF: Dollar amount of lifetime gifts to date

- LGIF: Dollar amount of largest gift to date
- RGIF: Dollar amount of most recent gift
- TDON: Number of months since last donation
- TLAG: Number of months between first and second gift
- AGIF: Average dollar amount of gifts to date
- DONR: Classification Response Variable (1 = Donor, 0 = Non-donor)
- DAMT: Prediction Response Variable (Donation Amount in $).

Note that the DONR and DAMT variables are set to "NA" for the test set. Use the guidelines provided in the R script file "CourseProjectEx.R" (available in Canvas) to fulfill the following requirements.

COURSE PROJECT REQUIREMENTS

1. Conduct exploratory data analysis on the data set prior to building classification and prediction models. In the report, do not devote a lot of space to discussions of dead-ends, pursuit of unproductive ideas, coding problems, etc.

2. Develop a classification model for the DONR variable using any of the variables as predictors (except ID and DAMT). Fit all candidate models (logistic regression, logistic regression GAM, LDA, QDA, *k*-nearest neighbors, etc.) using the training data and evaluate the fitted models using the validation data. Use "maximum profit" as the evaluation criteria and use your final selected classification model to classify DONR responses in the test dataset (the R script file "CourseProjectEx.R" provides some details).

3. Develop a prediction model for the DAMT variable using any of the variables as predictors (except ID and DONR). Fit all candidate models (least squares regression, best subset selection with *k*-fold cross-validation, principal components regression, partial least squares, ridge regression, lasso, etc.) using the training data and evaluate the fitted models using the validation data. Use "mean prediction error" as the evaluation criteria and use your final selected prediction model to predict DAMT responses in the test dataset (the R script file "CourseProjectEx.R" provides some details).

4. Save your test set classifications and predictions into a csv file and submit to Canvas by the project deadline (the R script file "CourseProjectEx.R" provides details for how to do this). Also, be sure to submit the project report (PDF file) and R code to Canvas by the project deadline.

5. You must write up your course project results in a professional report, which should be no more than 15 single-spaced pages long. The report should include substantive details of your analysis, and it should have several sections (e.g. Introduction, Analysis, Results, Conclusions). The report should provide sufficient details that anyone with a reasonable statistics background could understand exactly what you have done. You should consider using tables and figures to enhance your report. Do not include the R code in your report, as it will be submitted as a separate file.

COURSE PROJECT GRADING CRITERIA (300 POINTS)

- 75 points based on the key patterns and insights discovered from your exploratory data analysis.
- 100 points based on the number, appropriateness and uniqueness of the classification models employed, as well as the profit you achieve for your best classification model on the test set.
- 100 points based on the number, appropriateness and uniqueness of the prediction models employed, as well as the mean prediction error you achieve for your best prediction model on the test set.
- 25 points based on the quality of your report (including: adherence to report guidelines; clarity of writing; organization and layout; appropriate use of tables and figures; careful proof-reading to minimize typos, incorrect spelling and grammatical errors).

HINTS

1. Starting by running the code in the R script file "CourseProjectEx.R." Then adapt the code to build your own models. The script file includes LDA, logistic regression, and linear regression. However, you should apply as many of the machine learning techniques we've covered in class as you can, as well as methods not covered in the class (e.g., gradient boosting).

2. Feel free to use any transformations of the predictors variables – some are included in the R script file as example. However, **do not** transform either DONR or DAMT. The predictor transformations in the R script file are purely illustrative. You can use any transformations you can think of for any of the predictors (e.g., Box-Cox family of power transformations, indicator variables for certain "interesting" quantitative predictors, etc.).

3. It is worth spending some time seeing if there are any unimportant predictor terms that are merely adding noise to the predictions, thereby harming the ability of the model to predict test data. Simplifying your model by removing such terms can bring model improvements.

4. To calculate profit for a particular classification model applied to the validation data, remember that each donor donates $14.50 on average and each mailing costs $2.00. So, to find an "ordered profit function" (ordered from most likely donor to least likely):
    a. Calculate the posterior probabilities for the validation dataset.
    b. Sort DONR in order of the posterior probabilities from highest to lowest.
    c. Calculate the cumulative sum of (14.5 x DONR – 2) as you go down the list.
    d. Then, find the maximum of this profit function. The R script file "CourseProjectEx.R" describes how to do this.

5. To classify DONR responses in the test dataset, you need to account for the "weighted sampling" (sometimes called over-sampling). Since the validation data response rate is 0.5 but the test data response rate is 0.1, the optimal mailing rate in the validation data needs to be adjusted before your apply it to the test data. Suppose the optimal validation mailing rate (corresponding to the maximum profit) is 0.7:
    a. Adjust this mailing rate using $0.7/(0.5/0.1) = 0.14$.
    b. Adjust the "non-mailing rate" using $(1 – 0.7)/((1 – 0.5)/(1 – 0.1)) = 0.54$.

c. Scale the mailing rate so that it is a proportion: 0.14/(0.14 + 0.54) = 0.206.
d. The optima test mailing rate is thus 0.206. The R script file "CourseProjectEx.R" provides full details of how to do this adjustment.

6. Remember that this is a **group** project, so please be sure to collaborate appropriately!