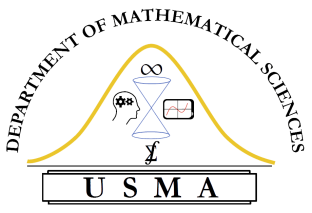
****

**MA478 – Generalized Linear Models**

Term End Analysis Competition (TEE)

**DESCRIPTION**

A charitable organization wishes to develop generalized linear models (GLMs) 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 GLM using data from the most recent campaign that can effectively capture 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 GLM 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\_full.csv.”

* ID: unique donor 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 within the “charity\_full.csv” file. Use the guidelines provided in the R script file “TEE\_Starter.R” to fulfill the following requirements.

**REQUIREMENTS**

1. Conduct thorough exploratory data analysis and appropriate data preparation procedures (transformations, feature engineering, etc.) on the data set prior to building GLM models.
2. Develop GLM models for the DAMT variable using any of the variables as predictors (except ID and DONR). Fit **at least three different** candidate models using the training data and evaluate the fitted models using the validation data. Use “mean squared error” as the evaluation criteria and use your final selected “best” GLM model to predict DAMT responses in the test dataset (the R script file “TEE\_Starter.R” provides details with an example linear regression model).
3. Save your test set classifications/predictions based upon your two “best” GLM models into a csv file (the R script file “TEE\_Starter.R” provides details for how to do this) and submit on Kaggle. Submit a PDF document of your well-documented statistical programming code on Teams.
4. Submit a summarized version of a statistical report on Canvas highlighting **data exploration** (including appropriate figures), **model**, and **analysis**. Your model section should include the statistical model with appropriate parameter estimates and standard errors. This should be written assuming the report will be read by the company’s data scientist who is well versed in statistics but perhaps not the intricacies of all Generalized Linear Models. Do not include an introduction, or conclusion. Assume that the data scientist has tasked you with this project so is well familiar with the problem and how to calculate profit.

**GRADING CRITERIA(200 POINTS)**

* 40 points based on the key patterns and insights discovered from your exploratory data analysis, along with the appropriateness and uniqueness of the data preparation procedures performed.
* 75 points based on the appropriateness and uniqueness of the GLM models built for prediction, as well as the mean absolute error you achieve for your “best” GLM-prediction model on the test set.
* 75 points based on the written analysis of the models including proper inference and key insights discovered and an overall analysis of expected profits from your ‘best’ model.
* 10 points based on the test set classification/prediction performance based upon your two “best” GLM models (evaluated using the “mean absolute error” metric on Kaggle; a lower score is better).

**HINTS**

1. Start by reviewing the starter code in the R script file “TEE\_Starter.R.” Then adapt the code to build your own GLM models. The script file includes example logistic regression and linear regression models. However, you should apply as many of the GLM techniques we’ve covered in class as you can, as well as statistical methods not covered in the class, where appropriate.
2. Feel free to use any transformations of the predictor variables – one is included in the R script file as example. However, **do not** transform either DONR or DAMT. The predictor transformation in the R script file is purely illustrative. You can use any transformations you can think of, where appropriate, 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 GLM 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):
   1. Calculate the posterior probabilities for the validation dataset.
   2. Sort DONR in order of the posterior probabilities from highest to lowest.
   3. Calculate the cumulative sum of (14.5 x DONR – 2) as you go down the list.
   4. Then, find the maximum of this profit function. The R script file “TEE\_Starter.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:
   1. Adjust this mailing rate using 0.7/(0.5/0.1) = 0.14.
   2. Adjust the “non-mailing rate” using (1 – 0.7)/((1 – 0.5)/(1 – 0.1)) = 0.54.
   3. Scale the mailing rate so that it is a proportion: 0.14/(0.14 + 0.54) = 0.206.
   4. The optima test mailing rate is thus 0.206. The R script file “TEE\_Starter.R” provides full details of how to do this adjustment.
6. Remember that this is an individual Term End Analysis Competition (TEE), so please be sure to work alone!

**URL:** <https://www.kaggle.com/t/adec9cf0884d30e7bbbe544c785c7933>