Stat897D – Applied Data Mining & Statistical Learning – Fall 2015

Team Project

This project is to be completed in your group – see the Course Schedule for the deadline.

A charitable organization wishes to develop a data-mining 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), the average donation is $14.50. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $2 to produce and send. Since expected profit from each mailing is 14.5 × 0.1 – 2 = –$0.55, it is not cost effective to mail everyone. We would like to develop a classification model 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 model to predict donation amounts for donors – the data for this will consist of the records for donors only. The data are available in the file "charity.csv" (available in Angel):

- 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 $).

The DONR and DAMT variables are set to "NA" for the test set. Use the guidelines provided in the R script file "TeamProjectEx.R" (available in Angel) to fulfill the following requirements.

Project Requirements

[Note: To help you with coding sample codes are provided in TeamProjectEx.R. You may want to modify it or write your own codes as per project requirement. You may not be able to use the code as is.]

1. Develop a classification model for the DONR variable using any of the variables as predictors (except ID and DAMT). Fit all candidate models 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 "TeamProjectEx.R" provides details).

2. Develop a prediction model for the DAMT variable using any of the variables as predictors (except ID and DONR). Use only the data records for which DONR=1. Fit all candidate models 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 "TeamProjectEx.R" provides details).

3. Save your test set classifications and predictions into a CSV file and one person from the team should submit this to Angel by the project deadline. Again, the R script file "TeamProjectEx.R" provides details for how to do this. Your test set classifications and predictions will be compared with the actual test set values of DONR and DAMT.

4. One person from the team should also submit the project report to Angel by the project deadline.

Write up your results in a professional report

- The report should be no more than 10 single-spaced pages long.
- It should include all substantive details of your analyses (the key word here is "substantive").
- The report should have sections (e.g., Introduction, Analysis, Results, Conclusion) and provide sufficient details that anyone with a reasonable statistics background could understand exactly what you've done.
- Feel free to *briefly* mention any exploratory aspects from your analyses, but do not devote a lot of space to discussions of dead-ends, pursuit of unproductive ideas, coding problems, etc.
- Consider using tables and figures to enhance your report.

- *Do not* embed R code in the body of your report; instead attach the code in an appendix. The appendix does *not* count towards the page limit.

Grading criteria (out of 25)

- 8 marks based on the profit you achieve for your classification model on the test set.
- 8 marks based on the mean prediction error you achieve for your prediction model on the test set.
- 9 marks for the quality of your report (including: clarity of writing, organization, and layout; appropriate use of tables and figures; careful proof-reading to *minimize* (not necessarily eliminate) typos, incorrect spelling, and grammatical errors; adherence to report guidelines above).

Hints

1. *Start by running the code in the R script file "TeamProjectEx.R." Then adapt the code to build your own models. The script file just includes linear discriminant analysis, logistic regression, and linear regression, but you should consider applying as many of the techniques we've covered in class as you can.*

2. Feel free to use any transformations of the predictor variables – some are included in the R script file as examples. 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:
   - Sometimes predictor transformations can be suggested by thinking about the underlying data. For example, one rationale for trying a quadratic transformation in a linear regression model is if you believe there is a possibility that the association between the response and that predictor (controlling for all the other included predictors) is non-linear. Perhaps average gifts tend to increase with donor's incomes but then level off when incomes are very high? Similar arguments can be made for classification models. Another type of transformation that may be suggested by the application at hand is indicator variables for certain "interesting" quantitative predictor variable values (e.g., if observations with X_1=0 behave differently to observations with X_1>0, then an indicator variable that is 1 for observations with X_1=0 and 0 for observations with X_1>0 may be helpful).
   - Other times, transformations are tried in a more exploratory, ad-hoc way. For example, a log transformation is often used for highly skewed variables (although sometimes the log transformation is "too strong" and a square root transformation may be better, while at other times the log transformation is "not strong enough" and a reciprocal or negative square root transformation may be better).
   - You can, if you wish, try either of these approaches for this project. You may not have enough time to be fully comprehensive in trying every possibility you can think of, so you may have to be selective and allocate your time carefully (just as in any real-world project where time and cost constraint always limit what you can do).

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. So, to find an "ordered profit function" (ordered from most likely donor to least likely):
   - Calculate the posterior probabilities for the validation dataset;
   - Sort DONR in order of the posterior probabilities from highest to lowest;

- Calculate the cumulative sum of (14.5 × DONR – 2) as you go down the list.

Then find the maximum of this profit function. The R script file "TeamProjectEx.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 you apply it to the test data. Suppose the optimal validation mailing rate (corresponding to the maximum profit) is 0.7:

    - Adjust this mailing rate using 0.7/(0.5/0.1) = 0.14;
    - Adjust the "non-mailing rate" using (1–0.7)/((1–0.5)/(1–0.1)) = 0.54;
    - Scale the mailing rate so that it is a proportion: 0.14/(0.14+0.54) = 0.206.

    The optimal test mailing rate is thus 0.206. The R script file "TeamProjectEx.R" provides full details of how to do this adjustment. Further details are available at http://blog.data-miners.com/2009/09/adjusting-for-oversampling.html. (If copying and pasting this link is unsuccessful, please type it in. It is a valid link.)

6. Remember that this is a team project, so work as a team. Everyone in the team will receive the same grade for the project unless it is brought to my attention that someone has tried to get a free ride, whereby his or her grade can be reduced.