ISE 5103 Intelligent Data Analytics
Homework #5

Instructor: Charles Nicholson

Due: See course website for due date

Learning objective: Perform classification modeling using logistic regression and tree-based techniques.
Submission notes:

1. You will submit exactly two files: a written document and your complete R script.
2. The write-up is your primary submission.
   - Clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.).
   - You **may** use "R Markdown" to *help* with your submission. Edit the final submission to *clearly and concisely* respond to the questions. **Limit this submission to under 10 pages**. Do not include large sections of code or code comments.
   - Failure to submit this file will automatically result in a grade penalty ($\geq 70\%$).
   - You will be graded primarily on your write-up.
3. You will also submit your R code.
   - *Provide comments* on what your code is doing. Keep it clean and clear!
   - Include `library` commands to load *all* packages that are used in the completion of the assignment at the beginning of your code.
   - Submissions without R code will automatically incur a grade penalty up to 30%.
4. Do not zip your files for submission. Name the files "LastName-HW1" with the appropriate file extension (that is, .R, .pdf, or .docx) – no HTML files.

1 Classification Performance Evaluation (16 points)

Create a user-defined R function that produces a series of model evaluations for a binary classifier. The function should produce at least the following items:

- Confusion matrix and statistics
- ROC curve and AUC
- Concordant Pairs
- D statistic
- K-S chart and statistic
- Distribution of predicted probabilities values for true positives and true negatives
- Lift chart (required!)

Note for this problem the only thing necessary to include in the write-up is the "documentation" for your function. That is, what is the name of the function, a description of the input parameters, and optionally, a description of the output values.

2 A data-driven approach to predict the success of bank telemarketing: Part 1 (14 points)

(a) Read the 2014 journal article by Moro et al., "A data-driven approach to predict the success of bank telemarketing" in Decision Support Systems, 62. The article is available in the course website.

(b) Write a one page review of the article.

3 A data-driven approach to predict the success of bank telemarketing: Part 2 (70 points)

The bank telemarketing data used in the above article is available at:
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
You will perform classification modeling on the same data. Please use the full dataset ($\sim$ 41k records).

(a) Carefully consider and explain your data splitting and/or cross-validation strategy.

(b) Using a logistic regression approach evaluate the influence, variance inflation, and residual diagnostics of your model.

(c) Additionally use elastic net regularization (for logistic regression), decision tree, and either random forest or a boosted tree to build the best classifier for test data. (No output is required for the write-up)

(d) Use your custom function from Problem #1 to help evaluate and compare the models developed. How do they compare with each other? How do you compare with Moro et al. (2014)?

(e) Which of your models do you recommend and why?

4 Extra-Credit (20 points)

(a) (10 points) Use a SVM as a classification model in the telemarketing problem; compare and contrast the results with the other classifiers.

(b) (10 points) Use a neural network as a classification tool in the telemarketing problem; compare and contrast the results with the other classifiers.