Homework 2

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

2017-01-18

Homework 2 I

You will use the dataset D8 described on DMR (online access via Pitt network) Appendix A for this assignment.

- Submit your report in PDF, and your code in *.R, via courseweb.
  - If you use “R Markdown,” make sure your code is reproducible, and you provide clear answers (elaborate description) to the questions.
- Due: 2017-01-31 11.59pm

Ref. DMR Appendix A

Homework 2 II

Task: analyze dataset D8 audit.csv
The objective is to predict the binary (TARGET_Adjusted) and continuous (RISK_Adjustment) target variables.

1. Read the data description on DMR p.297. Identify and report response variable and predictors.
2. Explore the dataset, and generate both statistical and graphical summary with respect to the numerical and categorical variables. Provide analysis similar to HW1 (2b)–(2e)..
   a) Generate a summary table for the data. For each numerical variable, list: variable name, mean, median, 1st quartile, 3rd quartile, and standard deviation.
   b) For numerical variables, plot the density distribution. Describe whether the variable has a normal distribution or certain type of skew distribution.

Homework 2 III

c) For each categorical predictor, generate the conditional histogram plot of response variable.

3. Apply logistic regression analysis to predict TARGET_Adjusted. Evaluate the models through cross-validation and on holdout samples. Interpret the effect of the predictors.

   a) Implement a 10-fold cross-validation scheme by splitting the data into training and testing sets. Use the training set to train a logistic regression model to predict the response variable. Examine the performance of different models by varing the number of predictors. Report the performance of the models on testing set using proper measures (accuracy, precision, recall, F1, AUC) and plots (ROC, lift).

   b) For the best model, compute the odds ratio and interpret the effect of each predictors.

Homework 2 IV

4. Apply linear and non-linear regression analysis to predict RISK_Adjustment. Evaluate the models through cross-validation and on holdout (leave-one-out or 10-fold cross-validation) samples. Provide details similar to HW1 (3).

hint: treat the given binary values “1/0” as “positive/negative”

hint: be careful about missing values in data