Homework 3

BUAN 6356

Read the instructions below before you submit your homework.

1. Use the following **file naming convention** for the file you upload on eLearning: *BUAN6356_Homework3_[Lastname and firstname initial].* For example, if your full name is Jerry Wayne Jones, then the file name should be ***BUAN6356_Homework3_JonesJ***.

2. **DO NOT** use an absolute directory path.

3. **DO NOT** change the file or variable names before importing it. If you rename the dataset name or any variable name, use your R script to do that.

4. Any assignment submitted after the deadline will be considered late and will not be graded.

Homework 3

A team collected data on email messages to create a classifier that can separate spam from non-spam email messages.

The dataset and short descriptions about the variables in the dataset are available from the following data archive: https://archive.ics.uci.edu/ml/datasets/spambase

1. Examine how each predictor differs between the spam and nonspam e-mails by comparing the spam-class average and nonspam-class average. Identify 10 predictors for which the difference between the spam-class average and non-spam class average is highest.

2. Partition the data into training and validation sets by allocating 80 percent of the observations to the training dataset and 20 percent of the observations to the validation dataset. Next, perform a linear discriminant analysis using the training dataset. Include only 10 predictors identified in the question above in the model.

3. Using the confusion matrix, lift chart, and decile chart for the validation dataset evaluate the effectiveness of the model in identifying spams.