Homework 2

BUAN 6356

Read the instructions below before you start your analysis.

1. Create a **R Markdown** document to submit your answers. Only one file should be submitted. That file should contain R code, R output and all the required explanations and answers to all questions.

2. **DO NOT** use an absolute directory path. I should be able to "knit" (run) your R Markdown document to a .html document without trying to find the input data in another directory. Test the "knit" process before uploading the document on eLearning. Assume that I have the .csv file mentioned below.

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

4. Label the charts appropriately. I should be able to figure out what information a chart is providing by looking at the chart and its labels (e.g., title, axis names, axis labels).

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

Homework 2

Ledoitte, a management consulting firm, is studying the roles played by experience and training in a system administrator's ability to complete a set of tasks in a specified amount of time. Ledoitte is interested in figuring out which administrators can complete given tasks within a specified time and those who are not.

Data are collected on the performance of 75 randomly selected administrators. They are stored in the file *SystemAdministrators.csv* (available on eLearning).

The variable *Experience* measures months of full-time system administrator experience, while *Training* measures the number of relevant training credits. The outcome variable *Completed* is either Yes or No, according to whether or not the administrator completed the tasks.

1. Using **ggplot2** package, create a **scatter plot** of *Experience* vs. *Training* using color or symbol to distinguish programmers who completed the task from those who did not complete it. Which predictor(s) appear(s) potentially useful for classifying task completion?

2. Run a **logistic regression** model with both predictors using the entire dataset as training data. Generate a **confusion matrix** and answer the following: among those who completed the task, what is the percentage of programmers incorrectly classified as failing to complete the task?

3. How much experience must be accumulated by a programmer with 6 training credits before his or her estimated probability of completing the task exceeds 0.6? *(Hint: in a logistic regression you can write the left hand-side of the regression equation as the log of odds).*