IE 425

Homework 3 (due May 15, Sunday @23:59)

1. Answer Question 4 on page 189 of the book "An Introduction to Statistical Learning with Applications in R", Second Edition, 2021.

2. Consider the data set "Default" in the package "ISLR", where "default" is the output attribute. Partition the data set into training and test sets with 75% going into the training set by using a seed value of 425. Using k-NN classification with different values of *k* between 1 and 10, determine the error rate, sensitivity, and specificity for the instances in the test set.

3. Consider the "Human Resources Analytics" problem which is based on the data set "HR.csv".Why are our best and most experienced employees leaving prematurely? Have fun with this database and try to predict which valuable employees will leave next. Fields in the dataset include:

- Satisfaction Level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Departments (column sales)
- Salary
- Whether the employee has left

Our goal is to predict the profile of employees who are likely to quit.

a) Partition the data set into training and test sets with 80% going into the training set by using a seed value of 425. Whenever you need to use set.seed function, use set.seed(425).

b) Fit a logistics regression model using the observations in the training dataset. Comment on which input attributes are important in making predictions.

c) Provide the Confusion Matrix along with sensitivity, specificity, precision and recall on the test set obtained by the logistic regression model.

d) Draw the ROC curve using the ROCR package and provide the auc value.