Statistical Methods Quiz #4 (50 minutes)

'25. 12. 17.(Wednesday) Cadet number(교번): Name: Score:

<Instructions>

1. It is open-book, closed-web test.
2. You are not allowed to discuss on the problems during the exam.
3. Import *customer.csv* data first.(Use all the string variables as factors.)

1. Import *customer.csv* dataset. Use the **str** function to examine the structure of the *customer* data. How many observations and how many variables does the dataset contain? [15 points]

<R-code>

<Output>

The number of observations:_____
The number of variables:_____

2. Randomly select 3500 observations from the *customer* data and save them as the training data(**customer.train**). Save the remaining observations as the test data(**customer.test**). [15 points]

<R-code>

3. Using the training data, fit a **classification tree** with **Segmentation** as the response variable and all other variables as explanatory variables. Also, visualize the fitted tree. [15 points]

<R-code>

<Output (plot)>

4. Using the test data, obtain the predicted response variable. Then, create a **confusion matrix** and compute the **accuracy**. [20 points]

<R-code>

<Output>

Accuracy:_____

5. Using the training data, fit a **random forest** model with **Segmentation** as the response variable and all other variables as explanatory variables. (Set the number of variables randomly selected at each split to 3.) Using the test data, obtain the predicted response variable. Then, create a **confusion matrix** and compute the **accuracy**. Based on accuracy, which model performs better: the **single decision tree** or the **random forest**? (Select one.) [20 points]

<R-code>

<Output>

Accuracy:_____

The (single decision tree / random forest) model has better performance.

6. For the random forest model, visualize the **variable importance**. Which variable is the most important? [15 points]

<R-code>

<Output (plot)>

The most important variable:_____