ACTSC 632 – Assignment 2 – due on July 24

This assignment will consider the German Credit data set available in the `CASdatasets` package in which it is named `credit`. As a team of data analysts for a credit card company, you have been asked to compare different classification methods and provide a recommendation to your manager for which one is the best to identify the "good" (i.e., creditworthy, no non-payments) vs the "bad" (not credit-worthy, having existing non-payments) credit files.

The goal of this assignment is to compare different classification methods based on trees for this problem and make a recommendation on what is the best method to use.

1. Determine the number of predictors in this data set, and whether they are quantitative of qualitative. How many observations are there? How many observations are classified as a "bad credit"? "good credit"?

2. To explore the data further, produce the following plots: (i) for each of the predictors age and duration, make a box plot showing the distribution of the observations, separately for the "good" and "bad" observations.; (ii) for the pairs duration & savings and duration & credit history, plot the observations (using duration on the x axis) and use different symbols for the "good" and "bad" observations. Comments on the plots you obtained.

3. To prepare data for the tree models, randomly split your data in 70% of the observations for training and 30% for testing.

4. Simply using recursive binary partitioning, obtain a tree for this classification problem.

   (a) How many leaves does your tree have?

   (b) How many factors were used to build this tree?

   (c) What is the deviance for this tree? (If you used something else than the default definition of deviance in R, please specify how is deviance determined).

   (d) Plot the tree you obtained using R. There should be enough information that given an observation, one could determine in which leaf it ends up.

   (e) Use your tree to make predictions for the test data set. Produce the confusion matrix corresponding to your tree and plot the ROC curve.

5. Now try to use pruning to see if you can improve your results.

   (a) Using the function `cv.tree` in R, determine the optimal level of complexity for the tree, i.e., the number of terminal nodes in the tree that minimizes the test prediction error (estimated by cross-validation). Does pruning the tree improve the deviance? The prediction error? Use the test set to answer the latter two questions.

   (b) If pruning helps, then produce the confusion matrix for the pruned tree based on the optimal level of complexity and plot the corresponding ROC curve.

6. Now try bagging and random forests to see if you can improve your results.

   (a) Provide the confusion table obtained using bagging.

   (b) Provide the confusion table obtained using random forests. How many variables were each split chosen from?

(c) According to the results obtained based on random forests, which predictors seem the most important? Provide data and/or plots to answer this question.

7. Conclude by making a suggestion as to which of the above methods is the best for this problem.

You are allowed to work in teams of 2. Please indicate your teammate in your submission. Your report should be submitted to the LEARN dropbox in R Markdown format that includes all code used. It should be professionally organized, with a brief introduction.