BT5152: Decision Making Technology for Business
Semester 1, AY 2018/19
Assignment 1 (10 Marks in total)

Due: 6pm of September 11, 2018

1. PROBLEM DESCRIPTION

The LendingClub is a peer-to-peer leading company that directly connects borrowers and potential lenders/investors. In this assignment, you will build classification models to predict whether or not a loan provided by LendingClub is likely to be a bad loan. In other words, you will use data from the LendingClub to predict whether a loan will be paid off in full or the loan will be charged off and possibly go into default.

2. DATASET DESCRIPTION

We will be using a subset of features (categorical and numeric) from the LendingClub website. The features we will be using are described in the code comments below.

1. 'grade', # grade of the loan
2. 'sub_grade', # sub-grade of the loan
3. 'short_emp', # one year or less of employment
4. 'emp_length_num', # number of years of employment
5. 'home_ownership', # home_ownership status: own, mortgage or rent
6. 'dti', # debt to income ratio
7. 'purpose', # the purpose of the loan
8. 'term', # the term of the loan
9. 'last_delinq_none', # has borrower had a delinquency
10. 'last_major_derog_none', # has borrower had 90 day or worse rating
11. 'revol_util', # percent of available credit being used
12. 'total_rec_late_fee', # total late fees received to day
13. target = 'bad_loans' # prediction target (y) (1 means risky, 0 means safe)

3. TASKS

1. (6 marks) Model the training data "loan_train.csv" using KNN, Naïve Bayes, C50 decision tree decision tree receptively. Report training accuracies and test accuracies on the training dataset "loan_train.csv" and test dataset "loan_test.csv" respectively.

   - Remember to scale your numerical variables properly and convert categorical variables by OneHot for KNN.

2. (6 marks) Now we practice rpart package. In order to avoid over fitting, prune the decision tree using three **pre-pruning** methods, and **post-pruning by best complexity parameter**. Compare the accuracies of fully-grown tree and 4 trees (both on training set and testing set) of the decision tree classifier. Discuss which tree gives you the best prediction results on the test set.

   - Before pruning (the fully-grown tree in this assignment), please set cp= 1e-05 (0.00001).
   - For the 3 pre-pruning, try minsplit = 800, minbucket = 200, and maxdepth = 3.
   - **This bullet is not a requirement for this assignment. You are encouraged to try other pre-pruning parameters or change cp before pruning to understand more about how pruning affect the accuracy on the training set and test set.**

4. Submissions and Grading

You need to submit two files. One is a *.R file and one is a *.PDF file of you results and answers. Name both files by your student number (e.g., A0123456X) and upload to IVLE workbin submission folder "A1".

In your R script you can assume that dataset files are in the same directory as the R script, i.e. `train_data <- read.csv("loan_train.csv", stringsAsFactors = TRUE)`

You may use any packages, but make sure your R file is runnable and has all the dependency packages imported `e.g. library(C50)` Any submission with an R file that's not executable will receive a fail grade.

The page limit of the pdf file is maximum 2 pages including everything. The formatting is A4, default margin, 12 font size, single-spacing. There is no need to try to fill 2 pages. Correct answers are much more important the length of your answers for grading.

You may revise and submit as many times before deadline. Make sure to remove any old version that you don't wish to be graded.

If you have questions about A1, feel free to email TA and cc me. Later, if you have questions about grading of A1, then you can email TA and cc me because TA (not me) will grade your assignment by following my grading rules listed below.

5. Grading Rules

1. Plagiarism is strictly prohibited.
2. For Q1, your R program should be correct and be executed properly to generate the same answers as the model answers of the TA.
3. For Q2, again you should have correct R code that can produce the model answer results before and after pruning.
4. TA can judge the quality of your code and deduct up to 2 marks. For example, if you included quite a number of unnecessary codes (which shows you do not really know which line of R command is the real one that helps you conduct analysis). You can provide comments into your codes to show your understanding.
5. TA can give you up to +2 bonus if you did an excellent job. The max of A1 is still 12 marks.