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 delinquincy
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**

More details to be announced. Roughly, it has been stated in TA’s slides during the tutorial.

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 on the next page.