FIT3152 Data analytics: Assignment 2

This assignment is worth 10% of your final marks in FIT3152.

Due: Saturday May 19th 2018

Note: Students are expected to work individually on this assignment.

How to submit: Submit your written report as a pdf file (.pdf).
Submit your R working as an R script (.R). Do not submit the data file.
Use the naming convention: Firstname.Lastname.studentID.{pdf,R}
Upload the two files to Moodle. Do not zip.

Objective:
The objective of this assignment is to gain familiarity with classification models using R.

You will be using a modified version of the German Credit Data, which contains information on 1000 customers, described by 20 decision attributes and a class attribute. Each instance is classified as 'Good credit risk' or 'Bad credit risk' (encoded as Class labels '1' and '2' respectively). Details of the original data, including the decision attributes follow the assignment description.

You are expected to use R for your analysis, and may use any R package. Set your R working directory to 'desktop', clear the workspace, set the number of significant digits to a sensible value, and use 'GCD' as the default data frame name for the whole data set. Read your data into R using the following code:

```
rm(list = ls())
options(digits=4)
#install any packages
GCD <- read.csv("GCD2018.csv")
```

We want to obtain a model that may be used to predict whether a new customer is at risk of defaulting a borrowed loan.

Assignment questions:

1. Explore the data: What is the proportion of 'Good credit risk' to 'Bad credit risk' cases? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-valued attributes. Is there anything noteworthy in the data?

2. Divide your data into a 70% training and 30% test set by adapting the following code (written for the iris data). Use your student ID as the random seed.

```
set.seed(XXXXXXXX) #random seed
train.row = sample(1:nrow(iris), 0.7*nrow(iris))
iris.train = iris[train.row,]
iris.test = iris[-train.row,]
```

3. Create a classification model using each of the following techniques. For this question you may use each of the R functions at their default settings, or with minor adjustments to set factors etc.

- Decision Tree
- Naïve Bayes
- Bagging
- Boosting
- Random Forest

4. Using the test data, classify each of the test cases as ‘Good credit risk’ or ‘Bad credit risk’. Create a confusion matrix and report the accuracy of each model.

5. Using the test data, calculate the confidence of predicting a ‘Good credit risk’ for each case and construct an ROC curve for each classifier. You should be able to plot all the curves on the same axis. Use a different colour for each classifier. Calculate the AUC for each classifier.

6. Create a table comparing the results in parts 4 and 5 for all classifiers. Is there a single “best” classifier?

7. Examining each of the models, determine the most important variables in predicting whether or not an applicant is a good or bad credit risk. Which variables could be omitted from the data with very little effect on performance? Give reasons.

8. By experimenting with parameter settings for at least one of the classifiers, create the best classifier you can – that is, one with an accuracy greater than the models you originally created in Part A. Demonstrate this improved accuracy using ROC, AUC, or other accuracy measures. Report the parameter settings and assumptions made in designing this classifier.

9. Using the insights from your analysis so far, implement an Artificial Neural Network classifier and report its performance. Comment on attributes used and your data pre-processing required. How does this classifier compare with the others? Can you give any reasons?

10. Write a brief report (4 pages max) summarizing your results in parts 1 – 9. Use commenting (# ----) in your R script, where appropriate, to help a reader understand your code.

Data description

Description of the German credit dataset.

The dataset classifies people described by a set of attributes as good or bad credit risks. It is desirable to identify people/customers who are more likely to default their loans, that is, who are at bad credit risk.

1. Title: German Credit Data

2. Source Information

[Obtained from UCI Machine Learning Repository]

[Originally attributed to as follows]

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

3. Number of Instances: 1000

The original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file "german_credit_risk.xls".

4. Number of Attributes german: 20 (7 numerical, 13 categorical)

5. Class labels: 1 => Good credit risk
   2 => Bad credit risk

6. Attribute description for german

Attribute 1: (qualitative)
Status of existing checking account
A11 : ... < 0 DM
A12 : 0 <= ... < 200 DM
A13 : ... >= 200 DM / salary assignments for at least 1 year

A14 : no checking account

Attribute 2: (numerical)
Duration in month

Attribute 3: (qualitative)
Credit history
A30 : no credits taken/ all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now

A33 : delay in paying off in the past
A34 : critical account/
other credits existing (not at this bank)

Attribute 4: (qualitative)
Purpose
A40 : car (new)
A41 : car (used)
A42 : furniture/equipment
A43 : radio/television
A44 : domestic appliances
A45 : repairs
A46 : education
A47 : (vacation - does not exist?)
A48 : retraining
A49 : business
A410 : others

Attribute 5: (numerical)
Credit amount

Attribute 6: (qualitative)
Savings account/bonds
A61 : ... < 100 DM
A62 : 100 <= ... < 500 DM
A63 : 500 <= ... < 1000 DM
A64 : .. >= 1000 DM
A65 : unknown/ no savings account

Attribute 7: (qualitative)
Present employment since
A71 : unemployed
A72 : ... < 1 year
A73 : 1 <= ... < 4 years
A74 : 4 <= ... < 7 years
A75 : .. >= 7 years

Attribute 8: (numerical)
Installment rate in percentage of disposable income

Attribute 9: (qualitative)
Personal status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Attribute 10: (qualitative)
Other debtors / guarantors
A101 : none
A102 : co-applicant
A103 : guarantor

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)

Property
A121 : real estate
A122 : if not A121 : building society savings agreement/ life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no property

Attribute 13: (numerical)
Age in years

Attribute 14: (qualitative)
Other installment plans
A141 : bank
A142 : stores
A143 : none

Attribute 15: (qualitative)
Housing
A151 : rent
A152 : own
A153 : for free

Attribute 16: (numerical)
Number of existing credits at this bank

Attribute 17: (qualitative)
Job
A171 : unemployed/ unskilled - non-resident
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/ highly qualified employee/ officer

Attribute 18: (numerical)
Number of people being liable to provide maintenance for

Attribute 19: (qualitative)
Telephone
A191 : none
A192 : yes, registered under the customers name

Attribute 20: (qualitative)
foreign worker
A201 : yes
A202 : no