ST4061 – Computer Intensive Statistical Analytics II
ST6041 – Statistical Analytics Implementations II
2018-2019
In-class test 1

NAME AND SURNAME: ……..

STUDENT NUMBER: ….……..

PROGRAM: ……

INTRUCTIONS

- Provide your answers in this document, after each question item.
- Paste the R code you used for each question item.
- **Save your files regularly.**

Your Word document will be copied directly from your account for assessment.

List of useful R packages and functions

Packages:

```
glmnet
mlbench
pROC
randomForest
```

Functions:

```
install.packages()
library()

as.matrix()
coef()
cv.glmnet()
cv.tree()
diag()
glmnet()
ncol()
nrow()
predict()
randomForest()
roc()
sample()
set.seed()
sum()
summary()
table()
tree()
varImpPlot()
```

Question

Load the following libraries and dataset into your R session, and split the dataset as follows:

```
library(mlbench)
data(Sonar)
N = nrow(Sonar)
P = ncol(Sonar)-1
M = 150
set.seed(1)
mdata = Sonar[sample(1:N),]
itrain = sample(1:N,M)
x = mdata[,-ncol(mdata)]
y = mdata$Class
xm = as.matrix(x)
```

The task with this dataset is to train statistical models to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock. The response variable of interest in this question is `Class`, which specifies the type of cylinder (either metal or rock) for each observation. All other variables in the dataset are used as potential predictors.

(1) How many observations are there in the test set?

(2) Use `cv.glmnet()` in order to optimise the LASSO for the training set. Quote the value of the optimal regularization parameter.

(3) Fit the LASSO to the training set, using the optimal value of the regularization parameter found in (2). If your implementation in (2) did not work, use a value of 0.01. Quote the coefficients of the fitted LASSO model for the intercept and the first 12 model parameters (you can copy-and-paste from the R window into this document).

(4) Fit a classification tree to the training set. How many variables were used in the tree? Name these variables.

(5) Fit a random forest to the training set and provide a plot of variable importance from that fit.

(6) Generate predictions from the classification tree and random forest that you obtained respectively in (4) and (5). Provide the confusion matrices for these predictions and quote the classification error rates for each model.

(7) Compute and compare the test-set AUC values for the tree and the random forest, using `pROC::roc()`. Which method is more accurate?