OLLSCOIL NA hÉIREANN, CORCAIGH
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

Examination Session and Year	Summer 2021
Module Code	ST4061 / ST6041
Module Title	Statistical Methods for Machine Learning II Machine Learning and Statistical Analytics II
Paper Number	1
External Examiner	Mr Andrew Maclaren
Head of the Department	Dr Kevin Hayes
Internal Examiners	Dr E. Wolsztynski
Instructions to Candidates	– Provide your answers in the answer document template provided (and not in this document). – **Save your files regularly.**
Duration of Paper	3 hours for completion + an extra 30 minutes for upload
Special Requirements	None

List of R packages required for this exam

caret
gbm
glmnet
ISLR
MASS
nnet
pROC
randomForest

List of useful R functions

- Analysis:

coef()
cor()
confusionMatrix()
cv.glmnet()
cbind()
data()
data.frame()
dim()
gbm()
glm()
glmnet()
mean()
model.matrix()
names()
nnet()
nrow()
predict()
randomForest()
rfe()
rfeControl()
roc()
round()
sample()
sqrt()
summary()
table()
train()
trainControl()
which()
varImp()

- Graphs:

legend()
par()
plot()

- Randomising:

sample()
set.seed()

Question 1 [25 marks]

No R coding is required for this question. Figure 1 below shows the classification decision boundaries obtained by training four different models on a 2D dataset (X1,X2,Y) where Y is a 2-class categorical response variable.

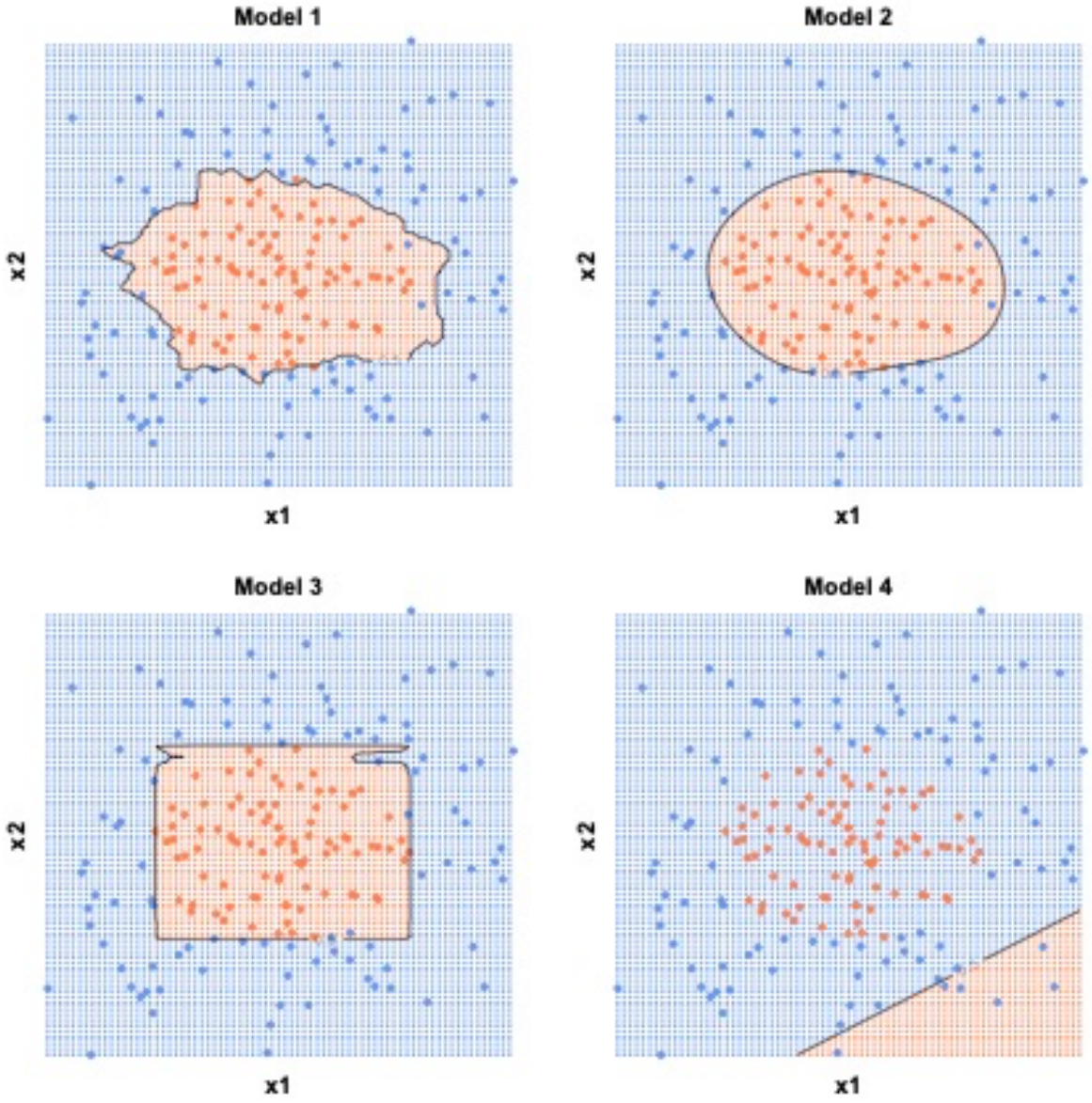

Figure 1 - Classification boundaries from 4 models on the same 2-class dataset (X1, X2, Y); blue and red dots depict the labelled data, and the areas correspond to the decision boundaries of each classifier.

(a) Indicate, in the table provided in your answer document, which of the scenarios represented in Figure 1 the models proposed in column A correspond to. (Indicate N/A if a model proposed in column A does not match any of the scenarios of Figure 1.) In column C, provide a brief explanation for your answer about each of the proposed models.

(b) Table 1 below provides a training dataset containing five observations, three predictors, and a 2-class categorical response variable Y taking values "HIGH" and "LOW". Assume we use this data to train a kNN classifier and that scaling is not required for this task.

(i) Indicate what is the predicted value of test point (0,0,0) with k=1? Justify your answer.

(ii) Indicate what is the predicted value of test point (0,0,0) with k=3? Justify your answer.

Note: your answers to question (b) should not contain more than 2 sentences each.

Table 1 - Training dataset for kNN classification.

Observation	X1	X2	X3	Y
1	0	2	0	HIGH
2	3	0	0	HIGH
3	0	1	3	LOW
4	0	1	2	HIGH
5	-1	0	1	LOW

Question 2 [45 marks]

Load the dodgysales.csv dataset into R as follows:

```
dat = read.csv(file="data/dodgysales.csv", stringsAsFactors=TRUE)
```

The response variable is Sales. Perform a *single* training-validation split of the data (i.e. without cross-validation) as follows:

```
n = nrow(dat)
set.seed(6041)
i.train = sample(1:n, floor(.7*n))
dat.train = dat[i.train,]
dat.validation = dat[-i.train,]
```

(a) Is this a regression or a classification problem?

(b) Quote the number P of predictors present in this dataset.

(c) Create a scaled copy dat.s of dataset dat, using min-max normalisation (apply this scaling to the response variable also).

(i) Quote the 5-number summaries of dat.s$Sales, dat.s$BudgOp and provide the frequency distribution table for dat.s$Training.

(ii) Split the scaled data dat.s into training and validation subsets as follows:

```
dat.s.train = dat.s[i.train,]
dat.s.validation = dat.s[-i.train,]
```

Fit two single-layer feed-forward neural networks, using respectively 3 and 8 neurons in the hidden layer. Use the nnet library to do this and set the random seed to 6041 (set.seed(6041)) before performing any model fit. Quote the corresponding training Mean Squared Errors (MSEs).

(iii) Generate predictions for the validation set dat.s.validation from each of the neural networks trained in (ii). Quote the corresponding validation MSEs.

(iv) Suggest an explanation for the difference between the training and validation errors for each of these neural networks.

(d) Set random seed to 6041 (set.seed(6041)) and fit a gradient boosting model to the training data dat.train, using package gbm, and using 100 weak learners for this ensemble. Quote the corresponding training and validation MSEs.

(e) Set random seed to 6041 (set.seed(6041)) and fit a generalized linear regression model to the training data dat.train. Quote the corresponding

training and validation MSEs.

(f) Set random seed to 6041 (`set.seed(6041)`) and fit a ridge regression model to the training data `dat.train`. Quote the corresponding training and validation MSEs.

(g) Compare and comment on the validation errors obtained from the neural networks and from ridge regression.

(h) Set random seed to 6041 (`set.seed(6041)`) and perform feature elimination with `caret::rfe` on the training data `dat.train`, based on random forest modelling, using 10-fold cross-validation. Report the final subset of variables selected.

(i) Set random seed to 6041 (`set.seed(6041)`) and perform feature elimination with `caret::rfe` on the training data `dat.train`, based on backward stepwise elimination in logistic regression and using 10-fold cross-validation. Report the final subset of variables selected.

(j) Indicate which feature subset from (h) and (i) you would use for analysis, and why.

Question 3 [30 marks]

Load the `BreastCancer` dataset from package `mlbench` and prepare it for analysis as follows:

```
library(mlbench)
data(BreastCancer)
dat = na.omit(BreastCancer)
dat$Id = NULL
set.seed(4061)
i.train = sample(1:nrow(dat), 600, replace=FALSE)
dat.train = dat[i.train,]
dat.validation = dat[-i.train,]
```

In this question, the response variable is `Class`. You are required to use the `caret` package for this question.

a) Set random seed to 4061 (`set.seed(4061)`) and fit a random forest model to the training set, performing a simple 10-fold cross-validation for training. Obtain predictions from this model for the validation set `dat.validation`.

   (i) Quote the number of variables used at each split.

   (ii) Provide the test set prediction accuracy achieved with this model.

b) Set random seed to 4061 (`set.seed(4061)`) and fit a support vector machine with a linear kernel to the training set, performing a simple 10-fold cross-validation for training. Obtain predictions from this model for the validation set `dat.validation`. Provide the test set prediction accuracy achieved with this model.

c) Set random seed to 4061 (`set.seed(4061)`) and fit a support vector machine with a radial basis kernel to the training set, performing a simple 10-fold cross-validation for training. Obtain predictions from this model for the validation set `dat.validation`. Provide the test set prediction accuracy achieved with this model.

d) Which which model is deemed better? Justify your answer.

e) Which 3 predictors seem to be the most important ones in predicting tumour class? Explain your answer.