ST4061 – Computer Intensive Statistical Analytics II

2020-2021

In-class test 2

NAME AND SURNAME: ……Amit Somnath Sambrekar

STUDENT NUMBER: ……..120220153

PROGRAM: ……MSc in DATA SCIENCE AND ANALYTICS.……

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.

Question 1

Figure 1 below shows the output of a random forest model fit to a sample of data points. This dataset comprises of 5 variables: Length, Width, Leaf, Curve and Age (Young; Intermediate; Mature).

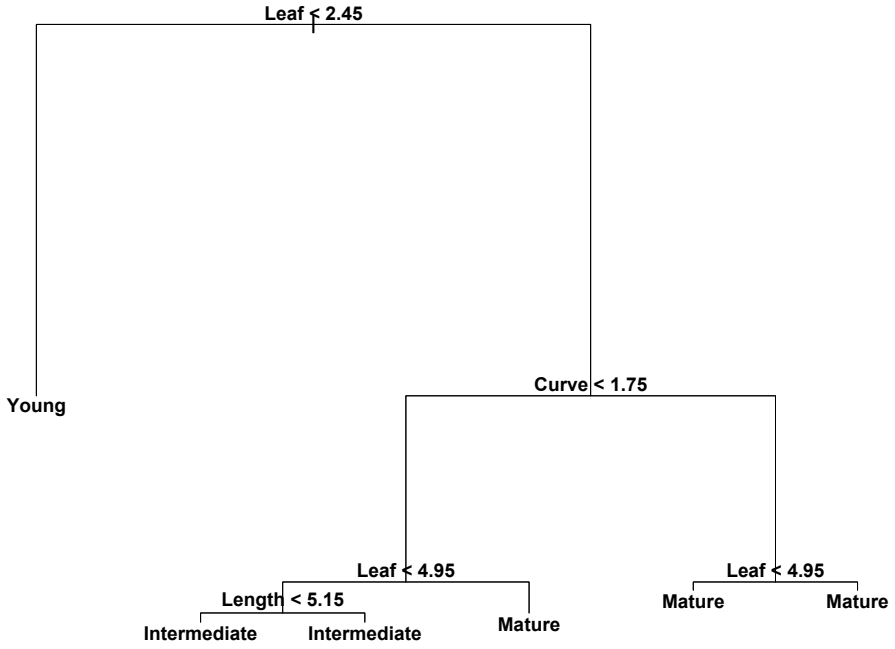


Figure 1 – Diagram for Question 1(a).

(1) Indicate which of the 5 variables is the response variable Y. Briefly justify your answer.

Answer to (1): Age is the response variable as it is the decision out variable in the random forest

(2) Quote the number of nodes that would be required in each layer of the architecture of a fully connected, single-hidden-layer, feed-forward neural network with 5 hidden neurons, for that model to be applicable to this dataset. Briefly justify your answer.

Answer to (2):

- **For the input layer, the number of nodes required would be 4 as this is number of predictors.**
- **The number of nodes in hidden layer would be 4.**
- **The number of nodes required in output layer would be 3, as the response variable is Age (Young; Intermediate; Mature).**

(3) How many model coefficients would need to be estimated in total, in order to fit the neural network described in (2) to this dataset? Briefly justify your answer.

Answer to (3): 43

Question 2

Run the following R instructions to load required libraries and create the dataset:

```
library(MASS)
library(tree)
library(randomForest)

dat = Boston   # NOTE: this is a data.frame...
set.seed(6041)
dat = dat[sample(1:nrow(dat)),]
dat$zn <- NULL
X = dat
X$medv <- NULL
X = scale(X)   # NOTE: this makes it a matrix...
Y = dat$medv
```

Here the response variable of interest Y, corresponds to the median value of owner-occupied homes in suburbs of Boston in $1,000's. All other variables in the dataset are used as potential predictors. Note that these predictors have been scaled.

Note: do **not** use the caret package for this question.

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

Answer to (1):12

(2) Calculate the *correlation* matrix of the scaled predictor set X.
   (a) Quote the first 3 x 3 elements of this matrix (i.e. its 3 x 3 top-left corner).
   (b) Inspect this matrix for correlations of magnitude greater than 90%, and comment on what action such occurrences may prompt (you are not asked to perform any further action). Justify your answer.

Answer to (2):a)

R code: a)

```
              crim       indus         chas
crim     1.00000000  0.40658341 -0.055891582
indus    0.40658341  1.00000000  0.062938027
chas    -0.05589158  0.06293803  1.00000000
```

b)

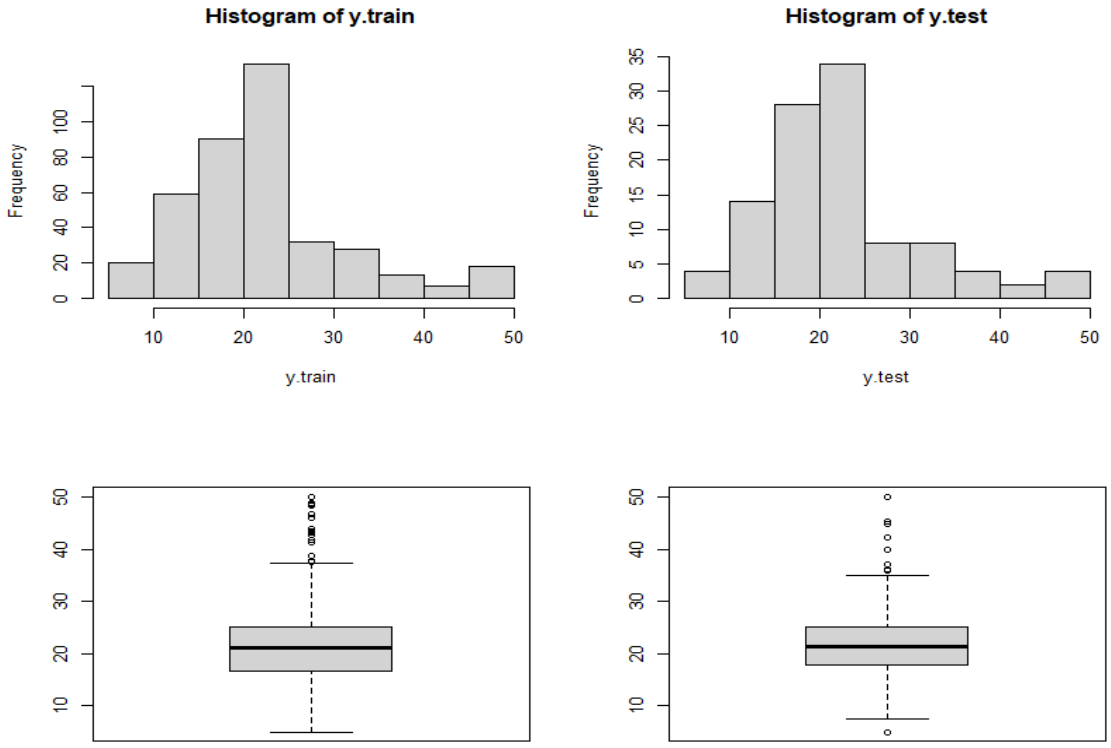
There is a linear relationship between the variable rad(index of accessibility to radial highways.) and tax(full-value property-tax rate per $10,000.)
As the rad increases, there is an increase in tax value.
As both the variable is explaining the same thing, one of the variables can be removed from the model and can be tested for further valuation.

Note: do **not** perform any action on these predictors at this point or in the rest of the question.

(3) Split the data into training and test sets, using the first 400 observations for training and the remainder for testing. Provide relevant statistical summaries of the distribution of Y_{train} and Y_{test}.

Answer to (3):

```
itrain <- 1:400
x.train <- X[itrain,]
x.test <- X[-itrain,]
y.train <- Y[itrain]
y.test <- Y[-itrain]
set.seed(6041)
par(mfrow=c(2,2))
hist(y.train)
hist(y.test)
boxplot(y.train)
boxplot(y.test)
```



(4) Set the random seed to 6041 (set.seed(6041)). Fit a random forest to the training data, using the randomForest package, and leaving all model parameters to default values.
   (a) Calculate and quote the Mean Square Error (MSE) for the training set.
   (b) Calculate and quote the Mean Square Error (MSE) for the test set.
   (c) Explain the difference between training MSE and test MSE observed for this model.

Answer to (4):

R code: rf.oxy = randomForest(x=x.train, y=y.train,
xtest=x.test, ytest=y.test)

MSE for training set: 11.77704
Mean Square Error (MSE) for the test set.: 9.98

Surprisingly, the train MSE is higher compared to test MSE, it may happen due to small sample size of training set.

```
Call:
 randomForest(x = x.train, y = y.train, xtest = x.test,
ytest = y.test)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 11.77704
                    % Var explained: 85.22
                       Test set MSE: 9.98
                    % Var explained: 90.16
```

(5) Fit a single regression tree to the training data, using the tree package, and leaving all model parameters to default values.
   (a) Calculate and quote the Mean Square Error (MSE) for the training set.
   (b) Calculate and quote the Mean Square Error (MSE) for the test set.
   (c) Explain the difference in test MSE between the random forest and tree-based regression.

Answer to (5):

R code:

x.train_df <- as.data.frame(x.train)
tree.out <- tree(y.train ~., data = x.train_df)

```
> set.seed(6041)
> x.train_df <- as.data.frame(x.train)
> tree.out <- tree(y.train ~., data = x.train_df)
> pred_t <- predict(tree.out, newdata = x.train_df)
> mse.train.tree <- mean((y.train - pred_t)^2)
> mse.train.tree
[1] 15.08148
> x.test.datafr <- as.data.frame(x.test)
> predicts <- predict(tr.out, newdata =x.test.datafr )
> mse.test.tree <- mean((y.test - predicts)^2)
> mse.test.tree
[1] 20.29177
```

There is a big difference in MSE between random forest and tree-based regression because random forest model consists od multiple tree based on boot aggregating process compared to single tree model