Yuen Chen Wing

~~Re-submission~~

Q2. Updated

RMSC 4002 Assignment 3 1st term 2018/2019

1155077996

We shall use the same dataset "credit.csv" in assignment 2.

see first submission

No marks deducted in Q2

Question 1 (CTREE)

(a) Read in credit.csv and save it in *d*. Use the last six digits of your birth date as the random seed, (For example, if your birth date is Dec. 10th, 1996, the random seed is 961210), randomly sample 580 records from *d* as the training dataset and save it in *d1*. Save the other records in *d2* as the testing dataset.

(b) Using the training dataset *d1*, build a classification tree of Result with other variables. Using the default option in *rpart()* probably gives a very complicated tree. Therefore we should add in the option *control=rpart.control(maxdepth=3)* inside the *rpart()* function. (See *help(rpart)* and *help(rpart.control)* for more details).

(c) Plot the tree with *use.n=T* and print the result. Write down the classification rules from the output. Compute the confidence, support and capture for each rule.

(d) Produce the classification table and compute the training error rate.

(e) Apply the classification rules in (d) using the testing dataset *d2* and produce the corresponding classification table. Compute the testing error rate, precision, recall and F1 score.

Question 2 (ANN)

(a) Using the same dataset *d1* and *d2* in question 1 (a) as the training dataset and testing dataset. Fit an improved version *ann()* function with size=7, linout=T, maxit=500 and try=25 and save the output to *ann7*.

(b) Repeat part (a) using size=8, 9 and 10. Save the result to ann8, ann9 and ann10 respectively.

(c) Compare the final value in *ann7*, *ann8*, *ann9* and *ann10*. Choose the best (smallest) one and produce the classification table for the training dataset *d1*. Compute the training error rate.

(d) Use the best ANN model in part (c), produce the classification table for the testing dataset *d2* and hence compute the testing error rate, precision, recall and F1 score.

(e) Compare and comment on these results obtained with the results in Question 1.

Submit the files: asg3-1.pdf, asg3-2.pdf, and asg3-1.R (or asg3-1.Rmd), asg3-2.R (or asg3-2.Rmd) containing answers and fully commented R commands via Blackboard on or before November 27, 2018 and put a hard copy in the drop-box on or before November 28, 2018.

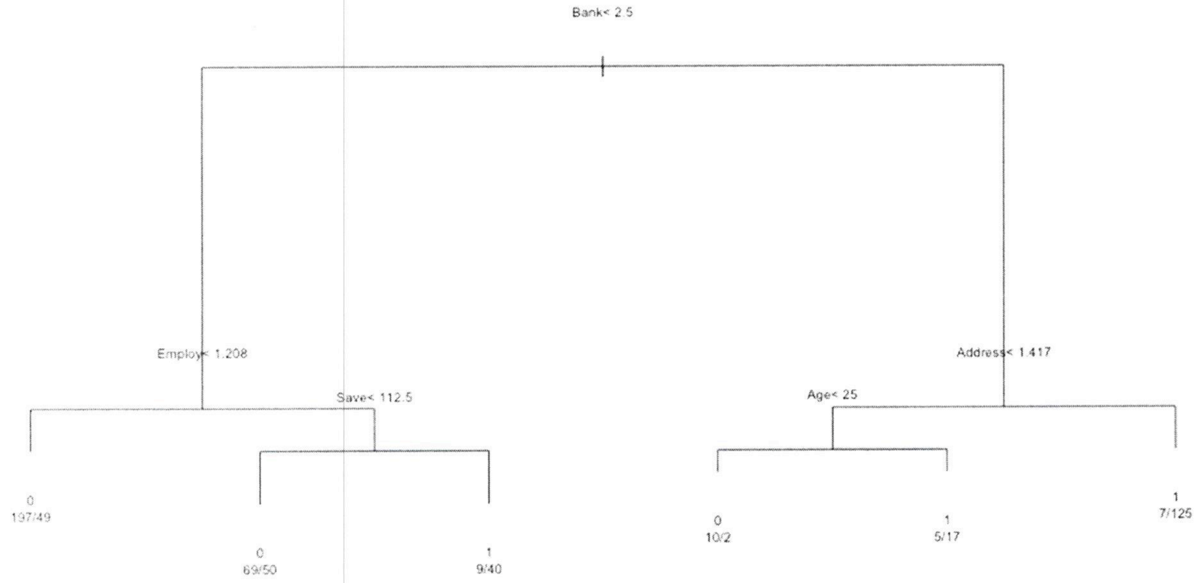
Reminder: Choose **one** member in your group to submit your final project on or before **December 4, 2018** and put a hard copy of your report in the drop-box on or before **December 5, 2018**.

Q1c. Plot the tree with use.n=T and print the result. Write down the classification rules from the output. Compute the confidence, support and capture for each rule.

Support = total number of cases in that rule / total number of cases in the entire dataset.

Confidence = proportion of correctly classified cases in that rule.
Capture = total number of cases of the majority group in that rule / total number of cases of that group

In Credit train data, there are 580 cases. Among them there are 283 observations of '1' and the remaining 297 belongs to '0' group.



R1 : If (Bank < 2.5) and (Employ < 1.208) then Result = 0 (197/49).
Support = (197+19)/580 = 0.3724, Confidence = 197/(197+49) = 0.8008, Capture = 197/283 = 0.6961.

R2 : If (Bank < 2.5) and (Employ > 1.208) and (Save < 112.5) then Result = 0 (69/50).
Support = (69+50)/580 = 0.2051, Confidence = 69/(69+50) = 0.5798, Capture = 69/(283) = 0.2438.

R3 : If (Bank < 2.5) and (Employ > 1.208) and (Save > 112.5) then Result = 1 (9/40).
Support = (9+40)/580 = 0.08448, Confidence = 40/(40+9) = 0.8163, Capture = 40/297 = 0.1347.

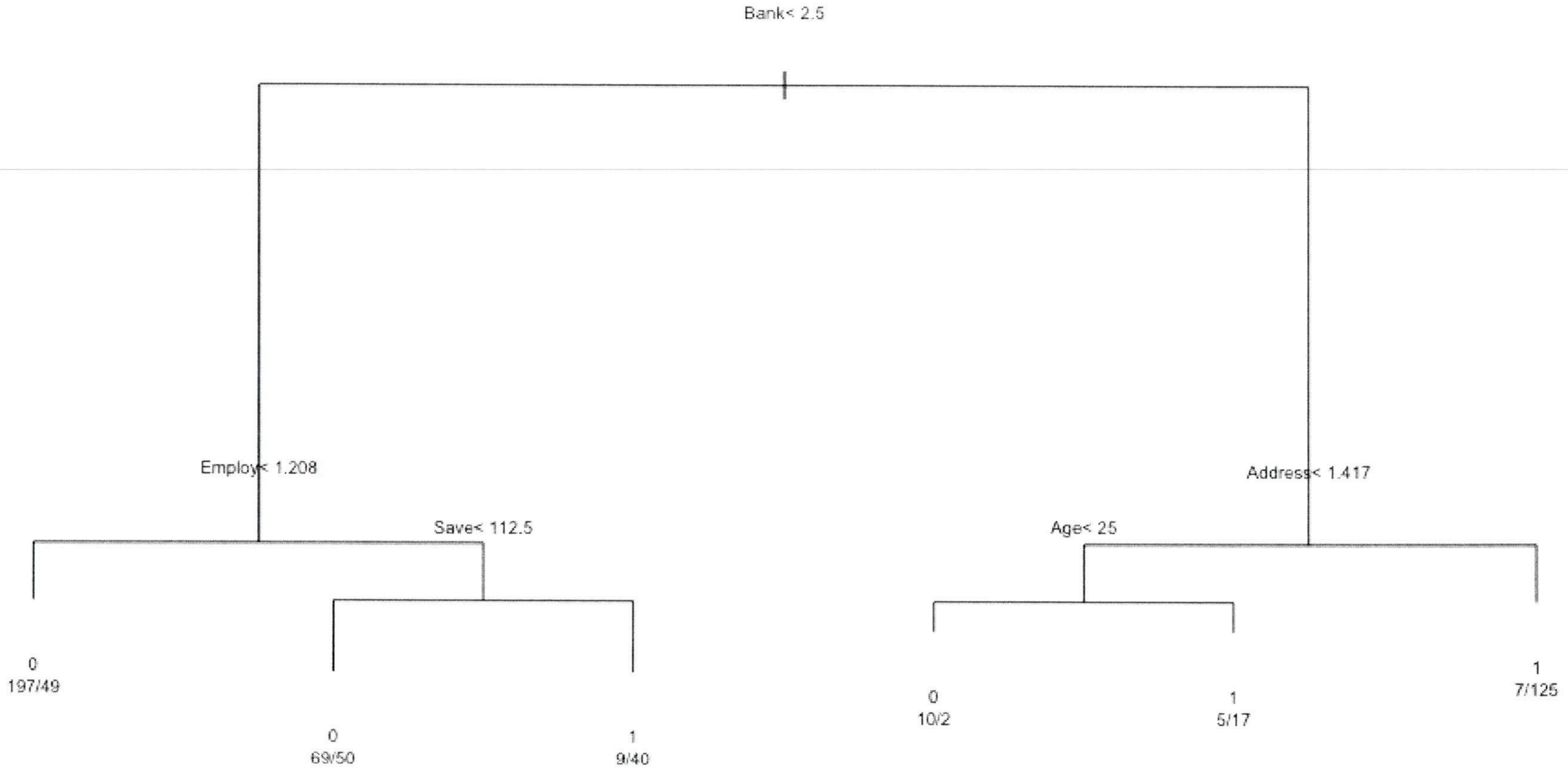
R4 : If (Bank > 2.5) and (Address < 1.417) and (Age < 25) then Result = 0 (10/2).
Support = (10+2)/580 = 0.02069, Confidence = 10/(10+2) = 0.8333, Capture = 10/283 = 0.03534.

R5 : If (Bank > 2.5) and (Address < 1.417) and (Age > 25) then Result = 1 (5/17).
Support = (5+17)/580 = 0.03793, Confidence = 17/(5+17) = 0.7727, Capture = 17/297 = 0..05723.

R6 : If (Bank > 2.5) and (Address <>1.417) then Result = 1 (7/125).

Support = (7+125)/580 = 0.2276, Confidence = 125/(125+7) = 0.9470, Capture = 125/297 = 0.4209.

Q1C


Bank< 2.5
Employ< 1.208
Save< 112.5
Address< 1.417
Age< 25
0
197/49
0
69/50
1
9/40
0
10/2
1
5/17
1
7/125

Q1d. Produce the classification table and compute the training error rate.

```
cl1  0  1

 1 276 101

 2  21 182
```

#the training error rate: (101+21)/580 = 0.2103448

Q1e. Apply the classification rules in (d) using the testing dataset d2 and produce the corresponding classification table. Compute the testing error rate, precision, recall and F1 score.

```
cl2  0  1

 1 79 13

 2  7 11
```

#the testing error rate: (13+7)/110 = 0.1818182

#Precsion Score: (79)/(79+13) = 0.8586957

#Recall Score: (79)/(79+7) = 0.9186047

#F1 Score: 2*0.8586957*0.9186047/(0.8586957+0.9186047) = 0.8876404

Q2c. Compare the final value in ann7, ann8, ann9 and ann10. Choose the best (smallest) one and produce the classification table for the training dataset d1. Compute the training error rate.

The final value in ann7-10 are 87.38877 101.36733 90.29165 84.89786 respectively.

Ann10 is selected, the classification table is shown as below:

cl1 0 1

0 248 57

1 49 226

#the training error rate: (60+49)/ 580 = 0.1827586

Q2d. Use the best ANN model in part (c), produce the classification table for the testing dataset d2 and hence compute the testing error rate, precision, recall and F1 score.

cl2 0 1

0 86 5

1 0 19

#the testing error rate: (5)/110 = 0.04545455

#Precsion Score: (86)/(86+5) = 0.9450549

#Recall Score: 85/(85) = 1

#F1 Score: 2*0.9450549*1/(0.9450549+1) = 0.9717514

Q2e. Compare and comment on these results obtained with the results in Question 1.

	Classification Tree	Artificial Neural Network
Test Error Rate	0.1818182	0.04545455
Precision	0.8586957	0.9450549
Recall	0.9186047	1
F1	0.8876404	0.9717514

Artificial neural network(Ann) has a superior performance than classification tree(Ctree). Ann in this question has lower test error rate and larger performance parameters than Ctree's. The possible reasons may be due to the model selection. We selected the best value among ann7-10 while we chose the Ctree with size=3 without any checking, there may exist better Ctree model. Another guess is that artificial neural network is more accurately fits the credit data.