Asg3-1.R

Leung Cheuk Wai Dominic 1155093086

Mon Nov 26 15:45:05 2018

```
#RMSC4002 HW3
#Leung Cheuk Wai Dominic 1155093086
#setwd("~/QFRM/RMSC4002/Assignment3") #For my own use only

#1a
d <- read.csv("credit.csv")
set.seed(980209) #set seed, my birth date is 9th Feb 1998
n <- nrow(d) #Get the length of the dataset
n #display length
```

```
## [1] 690
```

```
id <- sample(1:n,size=580)  #get the 580 random index for trianing data set
head(id) #display id
```

```
## [1] 178 689 371  12 100 653
```

```
d1 <- d[id,] #Save 580 data into training dataset d
dim(d1) #check dimension
```

```
## [1] 580   7
```

```
d2 <- d[-id,] #Save 110 data into testing dataset d1
dim(d2) #check dimension
```

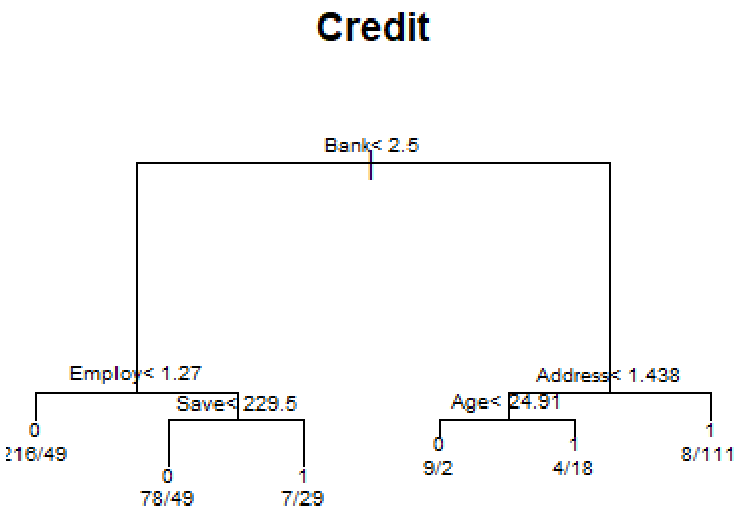
```
## [1] 110   7
```

```
#1b
library(rpart) #import library RPART
names(d) #Show the variables of credit.csv
```

```
## [1] "Age"     "Address" "Employ"  "Bank"    "House"   "Save"    "Result"
```

```
#Run the classification tree, added option control=rpart.control(maxdepth=3)
ctree <-
rpart(Result~Age+Address+Employ+Bank+House+Save,data=d1,method="class",contro
l=rpart.control(maxdepth = 3))

#1c
plot(ctree,asp=4,main="Credit") #Plot the branch of the tree
text(ctree,use.n=T,cex=0.6) #Add text to the tree
```



```
print(ctree) #Display the nodes

## n= 580
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 580 258 0 (0.55517241 0.44482759)
##    2) Bank< 2.5 428 127 0 (0.70327103 0.29672897)
##      4) Employ< 1.27 265  49 0 (0.81509434 0.18490566) *
##      5) Employ>=1.27 163  78 0 (0.52147239 0.47852761)
##       10) Save< 229.5 127  49 0 (0.61417323 0.38582677) *
##       11) Save>=229.5 36   7 1 (0.19444444 0.80555556) *
##    3) Bank>=2.5 152  21 1 (0.13815789 0.86184211)
##      6) Address< 1.4375 33  13 1 (0.39393939 0.60606061)
##       12) Age< 24.915 11   2 0 (0.81818182 0.18181818) *
##       13) Age>=24.915 22   4 1 (0.18181818 0.81818182) *
##      7) Address>=1.4375 119   8 1 (0.06722689 0.93277311) *

sum(d1["Result"]) #total of 1, accepted case

## [1] 258

nrow(d1) - sum(d1["Result"]) #total of 0, rejected case

## [1] 322
```

```
# For the person to be rejected, either rules below are satisfied:
#   1. If Bank<2.5 and Employ<1.27, then the person is rejected.
       # Support=(216+49)/580=0.4569, Confidence=216/265=0.8151,
Capture=216/322=0.6708
#   2. If Bank<2.5 and Employ>1.27 and Save<229.5, then the person is
rejected.
       # Support=(78+49)/580=0.2190, Confidence=78/127=0.6142,
Capture=78/322=0.2422
#   3. Bank>2.5 and Address<1.438 and Age<24.91, then the person is rejected.
       # Support=(9+2)/580=0.0190, Confidence=9/11=0.8181,
Capture=9/322=0.0280
# For the person to be accepted, either rules below are satisfied:
#   1. Bank<2.5 and Employ>1.27 and Save >229.5, then the person is accepted.
       # Support=(29+7)/580=0.0621, Confidence=29/36=0.8056,
Capture=29/258=0.1124
#   2. Bank>2.5 and Address<1.438 and Age>24.91, then the person is accepted.
       # Support=(18+4)/580=0.0379, Confidence=18/22=0.8181,
Capture=18/258=0.0698
#   3. Bank>2.5 and Address>1.438, then the person is accepted.
       # Support=(111+8)/580=0.205, Confidence=111/119=0.9328,
Capture=111/258=0.4302

#1d
pr <- predict(ctree) #Get probability of the sample
head(pr)
```

```
##             0         1
## 178 0.6141732 0.3858268
## 689 0.8150943 0.1849057
## 371 0.6141732 0.3858268
## 12  0.6141732 0.3858268
## 100 0.8150943 0.1849057
## 653 0.8150943 0.1849057
```

```
c1 <- max.col(pr) #Classify the sample with the larger probability;
1:rejected, 2:accepted
head(c1)
```

```
## [1] 1 1 1 1 1 1
```

```
table(c1,d1$Result) #Display the classification table
```

```
## 
## c1    0   1
##   1 303 100
##   2  19 158
```

```
#The Error rate = (100+19)/580=20.52%

#1e
```

```
pr2 <- predict(ctree,d2) #Get probability of testing data set
head(pr2)

##              0         1
## 3  0.19444444 0.8055556
## 4  0.06722689 0.9327731
## 11 0.61417323 0.3858268
## 13 0.81509434 0.1849057
## 17 0.18181818 0.8181818
## 19 0.81509434 0.1849057

c2 <- max.col(pr2) #Classify the sample with the larger probability
head(c2)

## [1] 2 2 1 1 2 1

table(c2,d2$Result) #Display the classification table

##
## c2   0  1
##   1 57 21
##   2  4 28

#The Error rate = (21+4)/110=22.73%
# Precision=28/(28+4)=85.5%
# Recall=28/(28+21)=57.14%
# F1 Score = 2/(1/Precision+1/Recall) = 69.13%
```