Chapter5_Example1_HSY.R

User

Sun Jan 13 21:46:08 2019

```
setwd("D:WWBITAminWWMachine Learning with R, Second Edition_CodeWWChapter 05")

Sys.setenv(JAVA_HOME="C:/Program Files/Java/jre1.8.0_191")
library(C50)
library(gmodels)
library(rJava)
library(RWeka)
library(party)
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'modeltools'
```

```
## The following object is masked from 'package:rJava':
##
##     clone
```

```
## Loading required package: strucchange
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
# 1
churn=read.csv("churn.csv", header=T, stringsAsFactors = T)
str(churn)
```

```
## 'data.frame':    5000 obs. of  20 variables:
##  $ state                         : Factor w/ 51 levels "AK","AL","AR",..: 12 27 36 33 41 13 2
9 19 25 44 ...
##  $ account_length                : int  101 137 103 99 108 117 63 94 138 128 ...
##  $ area_code                     : Factor w/ 3 levels "area_code_408",..: 3 3 1 2 2 2 2 1 3 2
...
##  $ international_plan            : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ voice_mail_plan               : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 2 ...
##  $ number_vmail_messages         : int  0 0 29 0 0 0 32 0 0 43 ...
##  $ total_day_minutes             : num  70.9 223.6 294.7 216.8 197.4 ...
##  $ total_day_calls               : int  123 86 95 123 78 85 124 97 117 100 ...
##  $ total_day_charge              : num  12.1 38 50.1 36.9 33.6 ...
##  $ total_eve_minutes             : num  212 245 237 126 124 ...
##  $ total_eve_calls               : int  73 139 105 88 101 68 125 112 46 89 ...
##  $ total_eve_charge              : num  18 20.8 20.2 10.7 10.5 ...
##  $ total_night_minutes           : num  236 94.2 300.3 220.6 204.5 ...
##  $ total_night_calls             : int  73 81 127 82 107 90 120 106 71 92 ...
##  $ total_night_charge            : num  10.62 4.24 13.51 9.93 9.2 ...
##  $ total_intl_minutes            : num  10.6 9.5 13.7 15.7 7.7 6.9 12.9 11.1 9.9 11.9 ...
##  $ total_intl_calls              : int  3 7 6 2 4 5 3 6 4 1 ...
##  $ total_intl_charge             : num  2.86 2.57 3.7 4.24 2.08 1.86 3.48 3 2.67 3.21 ...
##  $ number_customer_service_calls: int  3 0 1 1 2 1 1 0 2 0 ...
##  $ churn                         : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# 2
N=nrow(churn) ; N
```

```
## [1] 5000
```

```
set.seed(123)
sampling=sample(N, N*7/10)
head(sampling)
```

```
## [1] 1438 3941 2045 4413 4699  228
```

```
churn_train=churn[sampling, ]
churn_test=churn[-sampling, ]

# 3
prop.table(table(churn_train$churn))
```

```
##
##   no  yes
## 0.86 0.14
```

```
prop.table(table(churn_test$churn))
```

```
##
##        no       yes
## 0.8553333 0.1446667
```

```
# 4
churn_model= C5.0(churn_train[-20], churn_train$churn)
summary(churn_model)
```

```
##
## Call:
## C5.0.default(x = churn_train[-20], y = churn_train$churn)
##
##
## C5.0 [Release 2.07 GPL Edition]      Sun Jan 13 21:46:29 2019
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 3500 cases (20 attributes) from undefined.data
##
## Decision tree:
##
## total_day_minutes > 264.4:
## :...voice_mail_plan = yes: no (51/5)
## :   voice_mail_plan = no:
## :   :...total_eve_charge > 16.09: yes (99/4)
## :       total_eve_charge <= 16.09:
## :       :...total_day_minutes <= 277.5: no (26/3)
## :           total_day_minutes > 277.5:
## :           :...total_eve_minutes <= 138.5: no (14/3)
## :               total_eve_minutes > 138.5: yes (32/4)
## total_day_minutes <= 264.4:
## :...number_customer_service_calls > 3:
##     :...total_day_minutes > 160.2: no (158/33)
##     :   total_day_minutes <= 160.2:
##     :   :...total_eve_charge <= 19.76: yes (82/3)
##     :       total_eve_charge > 19.76:
##     :       :...total_day_minutes <= 138.4: yes (13/2)
##     :           total_day_minutes > 138.4: no (10)
##     number_customer_service_calls <= 3:
##     :...international_plan = yes:
##         :...total_intl_minutes > 13: yes (50)
##         :   total_intl_minutes <= 13:
##         :   :...total_intl_calls <= 2: yes (37)
##         :       total_intl_calls > 2: no (186/7)
##         international_plan = no:
##         :...total_day_minutes <= 220.8: no (2288/63)
##             total_day_minutes > 220.8:
##             :...total_eve_minutes > 242.3:
##                 :...voice_mail_plan = no: yes (63/20)
##                 :   voice_mail_plan = yes: no (21)
##                 total_eve_minutes <= 242.3:
##                 :...voice_mail_plan = yes: no (103/2)
##                     voice_mail_plan = no:
##                     :...total_eve_charge <= 17.47: no (204/12)
##                         total_eve_charge > 17.47:
##                         :...total_day_minutes <= 244.1: no (36/3)
##                             total_day_minutes > 244.1:
##                             :...total_night_minutes <= 214.4: no (13/2)
##                                 total_night_minutes > 214.4: yes (14)
##
##
## Evaluation on training data (3500 cases):
##
##      Decision Tree
```

```
##    ----------------
##    Size      Errors
##
##      20  166( 4.7%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##    2977    33    (a): class no
##     133   357    (b): class yes
##
##
##  Attribute usage:
##
##  100.00% total_day_minutes
##   93.66% number_customer_service_calls
##   86.14% international_plan
##   19.31% voice_mail_plan
##   15.51% total_eve_charge
##   14.29% total_eve_minutes
##    7.80% total_intl_minutes
##    6.37% total_intl_calls
##    0.77% total_night_minutes
##
##
## Time: 0.3 secs
```

```
# 5
churn_pred=predict(churn_model, churn_test)
CrossTable(churn_test$churn, churn_pred,
           prop.chisq = F, prop.c=F, prop.r=F,
           dnn=c('actual', 'predicted' ))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1500
##
##
##              | predicted
##       actual |        no |       yes | Row Total |
## -------------|-----------|-----------|-----------|
##           no |      1267 |        16 |      1283 |
##              |     0.845 |     0.011 |           |
## -------------|-----------|-----------|-----------|
##          yes |        56 |       161 |       217 |
##              |     0.037 |     0.107 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1323 |       177 |      1500 |
## -------------|-----------|-----------|-----------|
##
##
```

```
# 6
churn_boost10 = C5.0(churn_train[-20], churn_train$churn, trials=10)
churn_pred_boost10 = predict(churn_boost10, churn_test)
CrossTable(churn_test$churn, churn_pred_boost10,
           prop.chisq = F, prop.c=F, prop.r=F,
           dnn=c('actual', 'predicted' ))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1500
##
##
##              | predicted
##       actual |        no |       yes | Row Total |
## -------------|-----------|-----------|-----------|
##           no |      1275 |         8 |      1283 |
##              |     0.850 |     0.005 |           |
## -------------|-----------|-----------|-----------|
##          yes |        62 |       155 |       217 |
##              |     0.041 |     0.103 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1337 |       163 |      1500 |
## -------------|-----------|-----------|-----------|
##
##
```

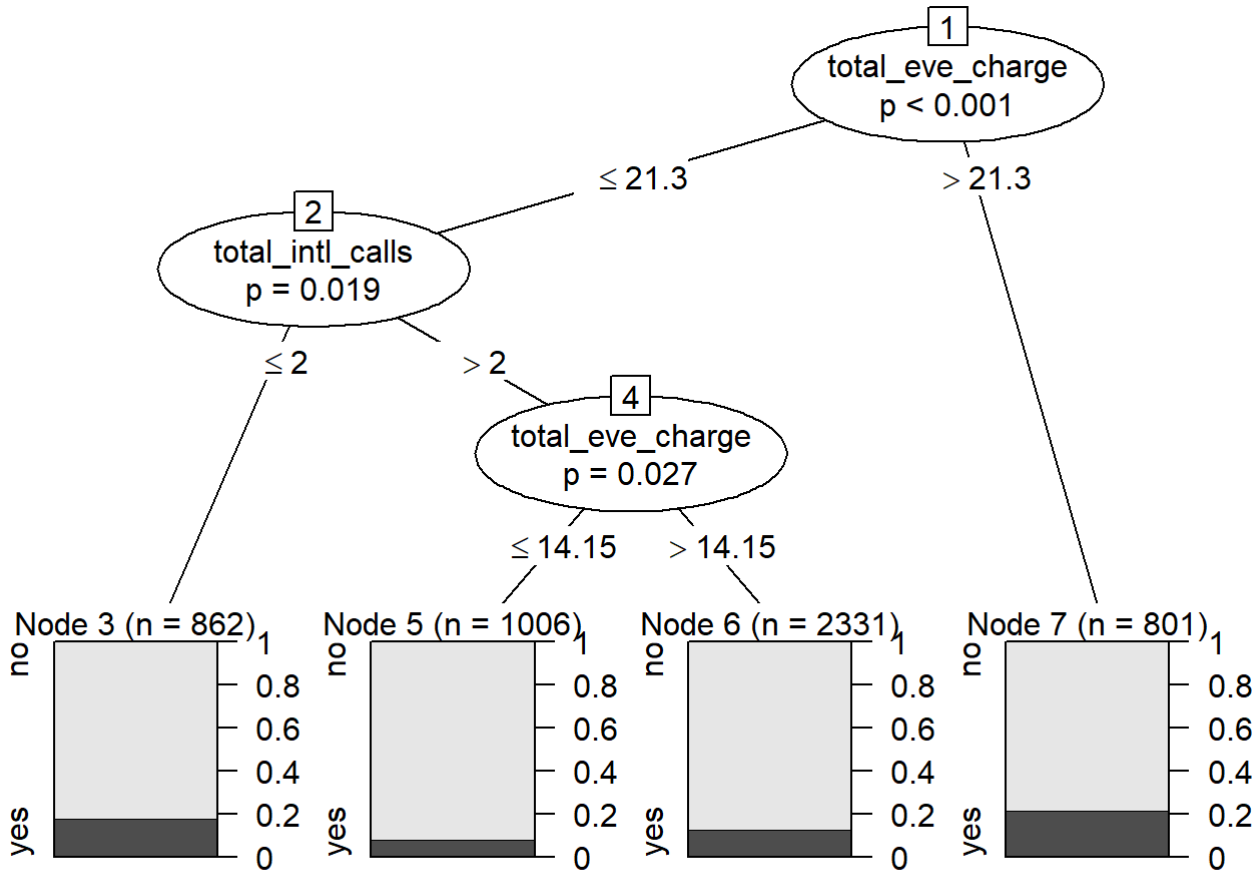
```
matrix_dimensions=list(c('no', 'yes'),c('no', 'yes'))
names(matrix_dimensions) =c('predicted', 'actual')
error_cost=matrix(c(0,3,1,0), nrow=2, dimnames=matrix_dimensions)
error_cost
```

```
##         actual
## predicted no yes
##       no   0   1
##       yes  3   0
```

```
churn_cost=C5.0(churn_train[-20], churn_train$churn,
                costs=error_cost)
churn_pred_cost=predict(churn_cost, churn_test)
CrossTable(churn_test$churn, churn_pred_cost, prop.chisq = F,
           prop.c=F, prop.r=F,
           dnn=c('actual', 'predicted'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1500
##
##
##              | predicted
##       actual |        no |       yes | Row Total |
## -------------|-----------|-----------|-----------|
##           no |      1280 |         3 |      1283 |
##              |     0.853 |     0.002 |           |
## -------------|-----------|-----------|-----------|
##          yes |        78 |       139 |       217 |
##              |     0.052 |     0.093 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1358 |       142 |      1500 |
## -------------|-----------|-----------|-----------|
##
##
```

```
# 7
churn_tree=ctree(churn~ total_intl_calls + total_night_calls
                 + total_day_calls + total_eve_charge, data=churn)
plot(churn_tree)
```


1
total_eve_charge
p < 0.001
≤ 21.3
> 21.3
2
total_intl_calls
p = 0.019
≤ 2
> 2
4
total_eve_charge
p = 0.027
≤ 14.15
> 14.15
Node 3 (n = 862)
Node 5 (n = 1006)
Node 6 (n = 2331)
Node 7 (n = 801)
no
yes
1
0.8
0.6
0.4
0.2
0