Lab07 - Arvores de Decisão

Machine Learning usando o R

Thalles Quinaglia Liduares

2022-08-12

Upload base de dados

```
setwd("C:\\Program Files\\R\\Dados\\ML")

data<-read.csv("credit.csv", stringsAsFactors = TRUE)

attach(data)
```

Upload pacotes

```
library(rpart)
library(rpart.plot)
library(caret)
library(rattle)
```

Analise exploratória dos dados

```
table(default)
```

```
## default
##  no yes
## 700 300
```

Indice de inadimplência igual a 30%.

```
summary(months_loan_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    12.0    18.0    20.9    24.0    72.0
```

O tempo minimo e maximo de duração dos empréstimos é igual a 4 e 72 meses, respectivamente. O tempo médio de duração dos empréstimos é igual a 21 meses.

```
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   27.00   33.00   35.55   42.00   75.00
```

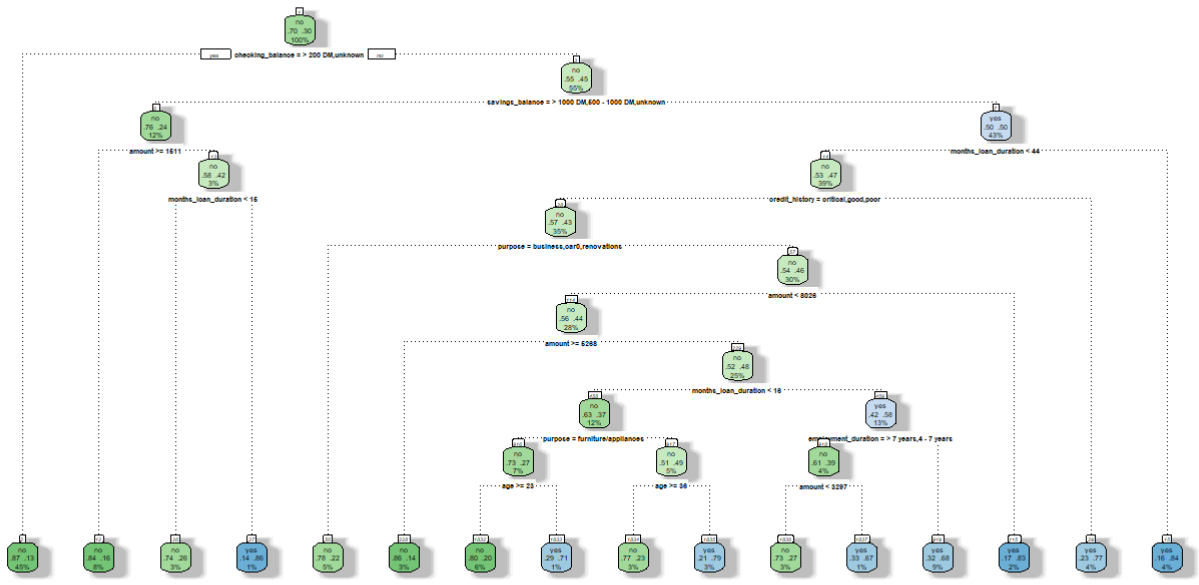
A idade média dos individuos clientes do banco é igual a 36 anos.

Divisão da amostra entre treino e teste.

```
set.seed(1608)

part_data<-floor(0.75*nrow(data))

treino_data <-sample(seq_len(nrow(data)), size = part_data)

treino<-data[treino_data, ]

teste<-data[-treino_data,]
```

Modelo

```
tree1<-rpart(default~., data = treino, method = "class")

fancyRpartPlot(tree1)
```



Rattle 2022-ago-12 17:37:31 USER

Validação

```
pred_tree1<-predict(tree1, newdata=teste, type="class")
```

Acurácia

```
teste$default<-as.factor(teste$default)

confusionMatrix(pred_tree1, teste$default)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no yes
##        no  150  38
##        yes  26  36
##
##                Accuracy : 0.744
##                  95% CI : (0.6852, 0.7969)
##     No Information Rate : 0.704
##     P-Value [Acc > NIR] : 0.09278
##
##                   Kappa : 0.3555
##
##  Mcnemar's Test P-Value : 0.16913
##
##             Sensitivity : 0.8523
##             Specificity : 0.4865
##          Pos Pred Value : 0.7979
##          Neg Pred Value : 0.5806
##              Prevalence : 0.7040
##          Detection Rate : 0.6000
##    Detection Prevalence : 0.7520
##       Balanced Accuracy : 0.6694
##
##        'Positive' Class : no
##
```

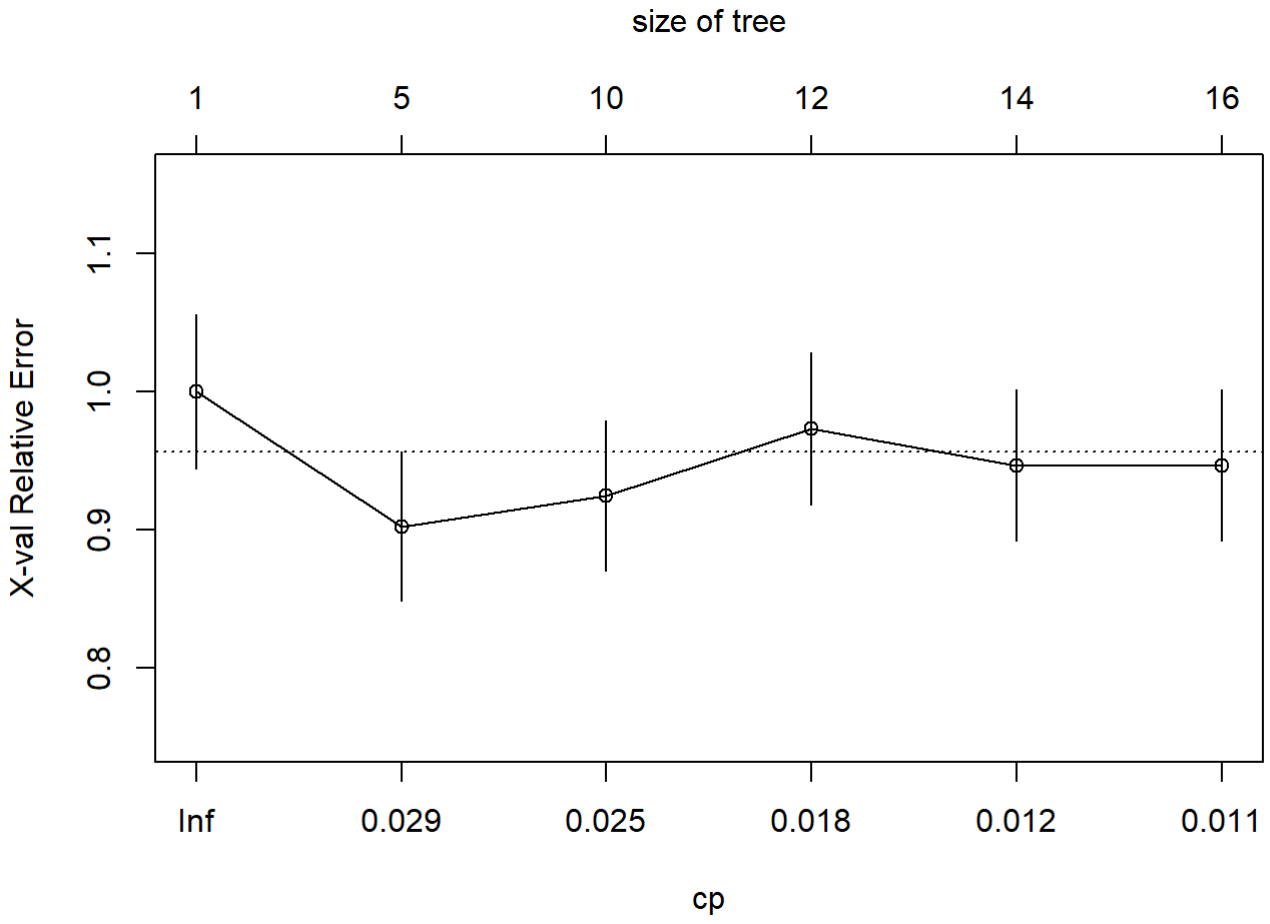
O nível de acurácia da árvore de decisão é igual a 74.4%.

Àrvore podada

```
printcp(tree1)
```

```
##
## Classification tree:
## rpart(formula = default ~ ., data = treino, method = "class")
##
## Variables actually used in tree construction:
## [1] age                 amount              checking_balance
## [4] credit_history      employment_duration  months_loan_duration
## [7] purpose             savings_balance
##
## Root node error: 226/750 = 0.30133
##
## n= 750
##
##         CP nsplit rel error  xerror     xstd
## 1 0.032448      0   1.00000 1.00000 0.055601
## 2 0.026549      4   0.82743 0.90265 0.053923
## 3 0.024336      9   0.69027 0.92478 0.054329
## 4 0.013274     11   0.64159 0.97345 0.055171
## 5 0.011062     13   0.61504 0.94690 0.054721
## 6 0.010000     15   0.59292 0.94690 0.054721
```

```
plotcp(tree1)
```




```
summary(tree1)
```

```
## Call:
## rpart(formula = default ~ ., data = treino, method = "class")
##   n= 750
##
##           CP nsplit rel error    xerror       xstd
## 1 0.03244838      0 1.0000000 1.0000000 0.05560077
## 2 0.02654867      4 0.8274336 0.9026549 0.05392279
## 3 0.02433628      9 0.6902655 0.9247788 0.05432912
## 4 0.01327434     11 0.6415929 0.9734513 0.05517092
## 5 0.01106195     13 0.6150442 0.9469027 0.05472052
## 6 0.01000000     15 0.5929204 0.9469027 0.05472052
##
## Variable importance
##     checking_balance months_loan_duration      savings_balance
##                   28                   14                   12
##               amount       credit_history                  age
##                   11                   11                    8
##              purpose  employment_duration existing_loans_count
##                    5                    5                    2
##   years_at_residence                  job
##                    1                    1
##
## Node number 1: 750 observations,    complexity param=0.03244838
##   predicted class=no   expected loss=0.3013333  P(node) =1
##     class counts:   524   226
##    probabilities: 0.699 0.301
##   left son=2 (340 obs) right son=3 (410 obs)
##   Primary splits:
##       checking_balance     splits as  RLRL,        improve=38.034780, (0 missing)
##       savings_balance      splits as  RLRLL,       improve=13.517980, (0 missing)
##       credit_history       splits as  LRRRR,       improve=13.018980, (0 missing)
##       months_loan_duration < 43.5    to the left,  improve= 8.657167, (0 missing)
##       housing              splits as  RLR,         improve= 5.862303, (0 missing)
##   Surrogate splits:
##       savings_balance      splits as  RLRLL,       agree=0.621, adj=0.165, (0 split)
##       credit_history       splits as  LRRRR,       agree=0.605, adj=0.129, (0 split)
##       employment_duration  splits as  RLRRR,       agree=0.560, adj=0.029, (0 split)
##       existing_loans_count < 1.5     to the right, agree=0.560, adj=0.029, (0 split)
##       months_loan_duration < 6.5     to the left,  agree=0.555, adj=0.018, (0 split)
##
## Node number 2: 340 observations
##   predicted class=no   expected loss=0.1264706  P(node) =0.4533333
##     class counts:   297    43
##    probabilities: 0.874 0.126
##
## Node number 3: 410 observations,    complexity param=0.03244838
##   predicted class=no   expected loss=0.4463415  P(node) =0.5466667
##     class counts:   227   183
##    probabilities: 0.554 0.446
##   left son=6 (88 obs) right son=7 (322 obs)
##   Primary splits:
##       savings_balance      splits as  RLRLL,       improve=9.667963, (0 missing)
##       months_loan_duration < 22.5    to the left,  improve=8.453034, (0 missing)
##       credit_history       splits as  LLRLR,       improve=7.846955, (0 missing)
##       amount               < 8732.5  to the left,  improve=4.491588, (0 missing)
```

```
##       housing              splits as  LLR,         improve=3.965200, (0 missing)
##   Surrogate splits:
##       age < 20.5    to the left,  agree=0.788, adj=0.011, (0 split)
##
## Node number 6: 88 observations,    complexity param=0.01106195
##   predicted class=no   expected loss=0.2386364  P(node) =0.1173333
##     class counts:    67    21
##    probabilities: 0.761 0.239
##   left son=12 (62 obs) right son=13 (26 obs)
##   Primary splits:
##       amount            < 1510.5  to the right, improve=2.510771, (0 missing)
##       credit_history    splits as  LRLLR,       improve=2.296320, (0 missing)
##       checking_balance  splits as  R-L-,        improve=2.049510, (0 missing)
##       percent_of_income < 3.5     to the left,  improve=1.533062, (0 missing)
##       purpose           splits as  LRRRR-,      improve=1.284965, (0 missing)
##   Surrogate splits:
##       months_loan_duration < 14.5    to the right, agree=0.773, adj=0.231, (0 split)
##       age                  < 59      to the left,  agree=0.750, adj=0.154, (0 split)
##       purpose              splits as  LLLRL-,      agree=0.739, adj=0.115, (0 split)
##
## Node number 7: 322 observations,    complexity param=0.03244838
##   predicted class=yes  expected loss=0.4968944  P(node) =0.4293333
##     class counts:   160   162
##    probabilities: 0.497 0.503
##   left son=14 (290 obs) right son=15 (32 obs)
##   Primary splits:
##       months_loan_duration < 43.5    to the left,  improve=8.245944, (0 missing)
##       credit_history       splits as  LLRLR,       improve=8.051945, (0 missing)
##       amount               < 8026    to the left,  improve=4.778301, (0 missing)
##       housing              splits as  LLR,         improve=3.992677, (0 missing)
##       age                  < 29.5    to the right, improve=2.592363, (0 missing)
##   Surrogate splits:
##       amount < 13501.5 to the left,  agree=0.91, adj=0.094, (0 split)
##
## Node number 12: 62 observations
##   predicted class=no   expected loss=0.1612903  P(node) =0.08266667
##     class counts:    52    10
##    probabilities: 0.839 0.161
##
## Node number 13: 26 observations,    complexity param=0.01106195
##   predicted class=no   expected loss=0.4230769  P(node) =0.03466667
##     class counts:    15    11
##    probabilities: 0.577 0.423
##   left son=26 (19 obs) right son=27 (7 obs)
##   Primary splits:
##       months_loan_duration < 14.5    to the left,  improve=3.609601, (0 missing)
##       years_at_residence   < 3.5     to the right, improve=2.219580, (0 missing)
##       age                  < 36.5    to the right, improve=2.219580, (0 missing)
##       percent_of_income    < 3.5     to the left,  improve=2.053419, (0 missing)
##       amount               < 1268.5  to the left,  improve=1.633484, (0 missing)
##   Surrogate splits:
##       purpose      splits as  LLRLL-,      agree=0.769, adj=0.143, (0 split)
##       amount       < 1219.5  to the left,  agree=0.769, adj=0.143, (0 split)
##       other_credit splits as  LLR,         agree=0.769, adj=0.143, (0 split)
##       job          splits as  RLLL,        agree=0.769, adj=0.143, (0 split)
##
```

```
## Node number 14: 290 observations,    complexity param=0.03244838
##   predicted class=no   expected loss=0.4655172  P(node) =0.3866667
##     class counts:   155   135
##    probabilities: 0.534 0.466
##   left son=28 (259 obs) right son=29 (31 obs)
##   Primary splits:
##       credit_history       splits as  LLRLR,       improve=6.614492, (0 missing)
##       months_loan_duration < 8.5     to the left,  improve=4.263725, (0 missing)
##       amount               < 625.5   to the left,  improve=3.656868, (0 missing)
##       housing              splits as  LLR,         improve=3.456722, (0 missing)
##       purpose              splits as  LRLRRL,      improve=2.689029, (0 missing)
##
## Node number 15: 32 observations
##   predicted class=yes  expected loss=0.15625  P(node) =0.04266667
##     class counts:     5    27
##    probabilities: 0.156 0.844
##
## Node number 26: 19 observations
##   predicted class=no   expected loss=0.2631579  P(node) =0.02533333
##     class counts:    14     5
##    probabilities: 0.737 0.263
##
## Node number 27: 7 observations
##   predicted class=yes  expected loss=0.1428571  P(node) =0.009333333
##     class counts:     1     6
##    probabilities: 0.143 0.857
##
## Node number 28: 259 observations,    complexity param=0.02654867
##   predicted class=no   expected loss=0.4285714  P(node) =0.3453333
##     class counts:   148   111
##    probabilities: 0.571 0.429
##   left son=56 (37 obs) right son=57 (222 obs)
##   Primary splits:
##       purpose              splits as  LRLRRL,      improve=3.893179, (0 missing)
##       amount               < 8026    to the left,  improve=3.523810, (0 missing)
##       months_loan_duration < 8.5     to the left,  improve=3.226572, (0 missing)
##       employment_duration  splits as  RRRLR,       improve=2.721712, (0 missing)
##       existing_loans_count < 2.5     to the right, improve=2.245898, (0 missing)
##
## Node number 29: 31 observations
##   predicted class=yes  expected loss=0.2258065  P(node) =0.04133333
##     class counts:     7    24
##    probabilities: 0.226 0.774
##
## Node number 56: 37 observations
##   predicted class=no   expected loss=0.2162162  P(node) =0.04933333
##     class counts:    29     8
##    probabilities: 0.784 0.216
##
## Node number 57: 222 observations,    complexity param=0.02654867
##   predicted class=no   expected loss=0.463964  P(node) =0.296
##     class counts:   119   103
##    probabilities: 0.536 0.464
##   left son=114 (210 obs) right son=115 (12 obs)
##   Primary splits:
##       amount               < 8026    to the left,  improve=3.461519, (0 missing)
```

```
##       months_loan_duration < 8.5     to the left,  improve=3.119950, (0 missing)
##       credit_history       splits as  LR-R-,       improve=2.410390, (0 missing)
##       age                  < 23.5    to the right, improve=2.251689, (0 missing)
##       employment_duration  splits as  RRRLR,       improve=2.172790, (0 missing)
##
## Node number 114: 210 observations,    complexity param=0.02654867
##   predicted class=no   expected loss=0.4428571  P(node) =0.28
##     class counts:   117    93
##    probabilities: 0.557 0.443
##   left son=228 (21 obs) right son=229 (189 obs)
##   Primary splits:
##       amount               < 5268    to the right, improve=4.200000, (0 missing)
##       months_loan_duration < 8.5     to the left,  improve=3.349074, (0 missing)
##       percent_of_income    < 2.5     to the left,  improve=3.033577, (0 missing)
##       age                  < 23.5    to the right, improve=2.478839, (0 missing)
##       employment_duration  splits as  RLRLR,       improve=2.100000, (0 missing)
##   Surrogate splits:
##       age < 69      to the right, agree=0.91, adj=0.095, (0 split)
##
## Node number 115: 12 observations
##   predicted class=yes  expected loss=0.1666667  P(node) =0.016
##     class counts:     2    10
##    probabilities: 0.167 0.833
##
## Node number 228: 21 observations
##   predicted class=no   expected loss=0.1428571  P(node) =0.028
##     class counts:    18     3
##    probabilities: 0.857 0.143
##
## Node number 229: 189 observations,    complexity param=0.02654867
##   predicted class=no   expected loss=0.4761905  P(node) =0.252
##     class counts:    99    90
##    probabilities: 0.524 0.476
##   left son=458 (93 obs) right son=459 (96 obs)
##   Primary splits:
##       months_loan_duration < 15.5    to the left,  improve=4.479263, (0 missing)
##       amount               < 653     to the left,  improve=3.467532, (0 missing)
##       employment_duration  splits as  RLRLR,       improve=2.333333, (0 missing)
##       age                  < 35.5    to the right, improve=2.092112, (0 missing)
##       credit_history       splits as  LL-R-,       improve=1.921308, (0 missing)
##   Surrogate splits:
##       amount              < 1543.5  to the left,  agree=0.709, adj=0.409, (0 split)
##       job                 splits as  RRRL,        agree=0.566, adj=0.118, (0 split)
##       employment_duration splits as  RLLRR,       agree=0.556, adj=0.097, (0 split)
##       age                 < 34.5    to the right, agree=0.556, adj=0.097, (0 split)
##       percent_of_income   < 3.5     to the left,  agree=0.545, adj=0.075, (0 split)
##
## Node number 458: 93 observations,    complexity param=0.02433628
##   predicted class=no   expected loss=0.3655914  P(node) =0.124
##     class counts:    59    34
##    probabilities: 0.634 0.366
##   left son=916 (52 obs) right son=917 (41 obs)
##   Primary splits:
##       purpose              splits as  -R-RL-,      improve=2.190442, (0 missing)
##       age                  < 23.5    to the right, improve=1.965426, (0 missing)
##       months_loan_duration < 8.5     to the left,  improve=1.946443, (0 missing)
```

```
##       amount               < 653     to the left,  improve=1.580749, (0 missing)
##       credit_history       splits as  LR-R-,       improve=1.312001, (0 missing)
##   Surrogate splits:
##       credit_history       splits as  RL-L-,       agree=0.667, adj=0.244, (0 split)
##       existing_loans_count < 1.5     to the left,  agree=0.645, adj=0.195, (0 split)
##       checking_balance     splits as  R-L-,        agree=0.634, adj=0.171, (0 split)
##       age                  < 32      to the left,  agree=0.613, adj=0.122, (0 split)
##       amount               < 2578    to the left,  agree=0.602, adj=0.098, (0 split)
##
## Node number 459: 96 observations,    complexity param=0.02654867
##   predicted class=yes  expected loss=0.4166667  P(node) =0.128
##     class counts:    40    56
##    probabilities: 0.417 0.583
##   left son=918 (31 obs) right son=919 (65 obs)
##   Primary splits:
##       employment_duration splits as  RLRLR,       improve=3.526220, (0 missing)
##       age                 < 54      to the right, improve=1.939394, (0 missing)
##       amount              < 4038.5  to the left,  improve=1.370410, (0 missing)
##       phone               splits as  RL,          improve=1.260417, (0 missing)
##       years_at_residence  < 1.5     to the left,  improve=1.195062, (0 missing)
##   Surrogate splits:
##       age                  < 35.5    to the right, agree=0.698, adj=0.065, (0 split)
##       months_loan_duration < 37.5    to the right, agree=0.688, adj=0.032, (0 split)
##
## Node number 916: 52 observations,    complexity param=0.01327434
##   predicted class=no   expected loss=0.2692308  P(node) =0.06933333
##     class counts:    38    14
##    probabilities: 0.731 0.269
##   left son=1832 (45 obs) right son=1833 (7 obs)
##   Primary splits:
##       age                  < 22.5    to the right, improve=3.2043960, (0 missing)
##       amount               < 708.5   to the left,  improve=1.3706290, (0 missing)
##       job                  splits as  RR-L,        improve=0.7786556, (0 missing)
##       months_loan_duration < 7.5     to the left,  improve=0.7091575, (0 missing)
##       existing_loans_count < 1.5     to the right, improve=0.7091575, (0 missing)
##   Surrogate splits:
##       savings_balance splits as  L-R--, agree=0.885, adj=0.143, (0 split)
##
## Node number 917: 41 observations,    complexity param=0.02433628
##   predicted class=no   expected loss=0.4878049  P(node) =0.05466667
##     class counts:    21    20
##    probabilities: 0.512 0.488
##   left son=1834 (22 obs) right son=1835 (19 obs)
##   Primary splits:
##       age                  < 35.5    to the right, improve=6.444743, (0 missing)
##       credit_history       splits as  LR-R-,       improve=4.526452, (0 missing)
##       years_at_residence   < 2.5     to the right, improve=1.724169, (0 missing)
##       months_loan_duration < 8.5     to the left,  improve=1.626694, (0 missing)
##       dependents           < 1.5     to the right, improve=1.626694, (0 missing)
##   Surrogate splits:
##       credit_history       splits as  LR-L-,       agree=0.732, adj=0.421, (0 split)
##       employment_duration  splits as  LLRLL,       agree=0.659, adj=0.263, (0 split)
##       years_at_residence   < 2.5     to the right, agree=0.659, adj=0.263, (0 split)
##       existing_loans_count < 1.5     to the right, agree=0.659, adj=0.263, (0 split)
##       months_loan_duration < 8.5     to the left,  agree=0.634, adj=0.211, (0 split)
##
```

```
## Node number 918: 31 observations,    complexity param=0.01327434
##   predicted class=no   expected loss=0.3870968  P(node) =0.04133333
##     class counts:    19    12
##    probabilities: 0.613 0.387
##   left son=1836 (22 obs) right son=1837 (9 obs)
##   Primary splits:
##       amount               < 3297    to the left,  improve=1.9824050, (0 missing)
##       age                  < 38.5    to the left,  improve=0.6508539, (0 missing)
##       months_loan_duration < 27      to the left,  improve=0.6144393, (0 missing)
##       housing              splits as  LLR,         improve=0.6144393, (0 missing)
##       existing_loans_count < 1.5     to the right, improve=0.2823270, (0 missing)
##   Surrogate splits:
##       months_loan_duration < 37.5    to the left,  agree=0.774, adj=0.222, (0 split)
##       credit_history       splits as  LL-R-,       agree=0.742, adj=0.111, (0 split)
##
## Node number 919: 65 observations
##   predicted class=yes  expected loss=0.3230769  P(node) =0.08666667
##     class counts:    21    44
##    probabilities: 0.323 0.677
##
## Node number 1832: 45 observations
##   predicted class=no   expected loss=0.2  P(node) =0.06
##     class counts:    36     9
##    probabilities: 0.800 0.200
##
## Node number 1833: 7 observations
##   predicted class=yes  expected loss=0.2857143  P(node) =0.009333333
##     class counts:     2     5
##    probabilities: 0.286 0.714
##
## Node number 1834: 22 observations
##   predicted class=no   expected loss=0.2272727  P(node) =0.02933333
##     class counts:    17     5
##    probabilities: 0.773 0.227
##
## Node number 1835: 19 observations
##   predicted class=yes  expected loss=0.2105263  P(node) =0.02533333
##     class counts:     4    15
##    probabilities: 0.211 0.789
##
## Node number 1836: 22 observations
##   predicted class=no   expected loss=0.2727273  P(node) =0.02933333
##     class counts:    16     6
##    probabilities: 0.727 0.273
##
## Node number 1837: 9 observations
##   predicted class=yes  expected loss=0.3333333  P(node) =0.012
##     class counts:     3     6
##    probabilities: 0.333 0.667
```

Acurácia com a àrvore podada

```
pod_tree <- prune(tree1,cp=tree1$cptable[which.min(tree1$cptable[,"xerror"]),"CP"])
pod_tree_fit <- predict(pod_tree,teste,type="class")
table(pod_tree_fit,teste$default)
```

```
##
## pod_tree_fit  no yes
##          no  170  57
##          yes   6  17
```

```
mean(pod_tree_fit==teste$default)
```

```
## [1] 0.748
```

Neste caso, o nivel de acurácia elevou-se para 74,8%.