Analysis

Elodie Kwan and Katia Voltz

2022-04-26

Import libraries and data

```
library(rpart)
German_credit <- read.csv("./../Data_DA/GermanCredit.csv", header = TRUE, sep = ";")

# German_credit$DURATION <- as.numeric(German_credit$DURATION)
# German_credit$AMOUNT <- as.numeric(German_credit$AMOUNT)
# German_credit$INSTALL_RATE <- as.numeric(German_credit$INSTALL_RATE)
# German_credit$AGE <- as.numeric(German_credit$AGE)
# German_credit$NUM_CREDITS <- as.numeric(German_credit$NUM_CREDITS)
# German_credit$NUM_DEPENDENTS <- as.numeric(German_credit$NUM_DEPENDENTS)
#
# for (i in 1:ncol(German_credit)){
#   if (class(German_credit[,i])=="integer"){
#     German_credit[,i] <- factor(German_credit[,i])
#     }
# }
```

Fitting a model :

Let's try a lassification tree

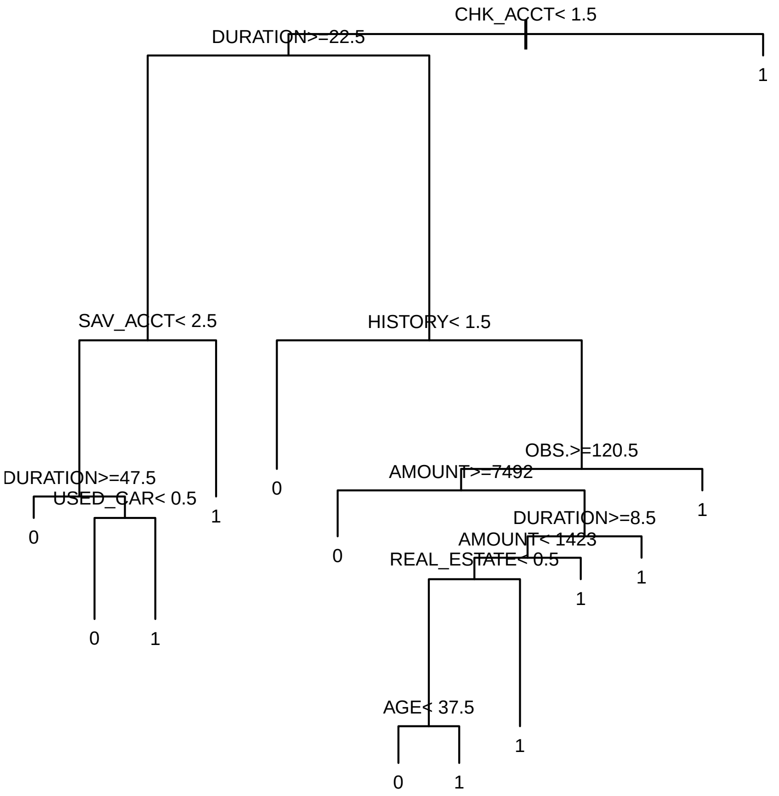
```
german.ct <- rpart(RESPONSE ~ ., method = "class", data = German_credit)
summary(german.ct)
```

```
## Call:
## rpart(formula = RESPONSE ~ ., data = German_credit, method = "class")
##   n= 1000
##
##           CP nsplit rel error    xerror       xstd
## 1 0.05166667      0 1.0000000 1.0000000 0.04830459
## 2 0.04666667      3 0.8400000 1.0066667 0.04839605
## 3 0.01833333      4 0.7933333 0.8800000 0.04646432
## 4 0.01400000      6 0.7566667 0.8600000 0.04612013
## 5 0.01333333     11 0.6866667 0.8633333 0.04617828
## 6 0.01000000     12 0.6733333 0.8966667 0.04674268
##
## Variable importance
##        CHK_ACCT        DURATION          AMOUNT         HISTORY        SAV_ACCT
##              30              14              10              10               9
##     REAL_ESTATE        USED_CAR            OBS.             AGE        RADIO.TV
##               5               4               4               3               3
##             JOB  PROP_UNKN_NONE       GUARANTOR MALE_MAR_or_WID      EMPLOYMENT
##               2               1               1               1               1
##     INSTALL_RATE
##               1
##
## Node number 1: 1000 observations,    complexity param=0.05166667
##   predicted class=1  expected loss=0.3  P(node) =1
##     class counts:   300   700
##    probabilities: 0.300 0.700
##   left son=2 (543 obs) right son=3 (457 obs)
##   Primary splits:
##       CHK_ACCT < 1.5     to the left,  improve=47.90962, (0 missing)
##       HISTORY  < 1.5     to the left,  improve=17.06212, (0 missing)
##       SAV_ACCT < 1.5     to the left,  improve=14.80642, (0 missing)
##       DURATION < 34.5    to the right, improve=13.62155, (0 missing)
##       AMOUNT   < 3913.5  to the right, improve=11.32017, (0 missing)
##   Surrogate splits:
##       SAV_ACCT   < 1.5     to the left,  agree=0.611, adj=0.149, (0 split)
##       HISTORY    < 3.5     to the left,  agree=0.592, adj=0.107, (0 split)
##       RADIO.TV   < 0.5     to the left,  agree=0.565, adj=0.048, (0 split)
##       EMPLOYMENT < 3.5     to the left,  agree=0.554, adj=0.024, (0 split)
##       AGE        < 30.5    to the left,  agree=0.554, adj=0.024, (0 split)
##
## Node number 2: 543 observations,    complexity param=0.05166667
##   predicted class=1  expected loss=0.441989  P(node) =0.543
##     class counts:   240   303
##    probabilities: 0.442 0.558
##   left son=4 (237 obs) right son=5 (306 obs)
##   Primary splits:
##       DURATION    < 22.5    to the right, improve=12.810640, (0 missing)
##       HISTORY     < 1.5     to the left,  improve= 9.653787, (0 missing)
##       REAL_ESTATE < 0.5     to the left,  improve= 9.181363, (0 missing)
##       SAV_ACCT    < 1.5     to the left,  improve= 8.890786, (0 missing)
##       AMOUNT      < 8079    to the right, improve= 6.601270, (0 missing)
##   Surrogate splits:
##       AMOUNT         < 2805.5  to the right, agree=0.748, adj=0.422, (0 split)
##       PROP_UNKN_NONE < 0.5     to the right, agree=0.643, adj=0.181, (0 split)
##       USED_CAR       < 0.5     to the right, agree=0.599, adj=0.080, (0 split)
##       REAL_ESTATE    < 0.5     to the left,  agree=0.597, adj=0.076, (0 split)
##       JOB            < 2.5     to the right, agree=0.595, adj=0.072, (0 split)
##
## Node number 3: 457 observations
##   predicted class=1  expected loss=0.131291  P(node) =0.457
##     class counts:    60   397
##    probabilities: 0.131 0.869
##
## Node number 4: 237 observations,    complexity param=0.05166667
##   predicted class=0  expected loss=0.4345992  P(node) =0.237
##     class counts:   134   103
##    probabilities: 0.565 0.435
##   left son=8 (196 obs) right son=9 (41 obs)
##   Primary splits:
##       SAV_ACCT     < 2.5     to the left,  improve=7.374515, (0 missing)
##       USED_CAR     < 0.5     to the left,  improve=4.129437, (0 missing)
##       AMOUNT       < 1381.5  to the left,  improve=3.289316, (0 missing)
##       INSTALL_RATE < 2.5     to the right, improve=3.067516, (0 missing)
##       DURATION     < 43.5    to the right, improve=2.564920, (0 missing)
## 
## Node number 5: 306 observations,    complexity param=0.04666667
##   predicted class=1  expected loss=0.3464052  P(node) =0.306
##     class counts:   106   200
##    probabilities: 0.346 0.654 
##   left son=10 (28 obs) right son=11 (278 obs)
##   Primary splits:
##       HISTORY     < 1.5     to the left,  improve=10.040510, (0 missing)
##       OBS.        < 120.5   to the right, improve= 6.207418, (0 missing)
##       REAL_ESTATE < 0.5     to the left,  improve= 5.585685, (0 missing)
##       GUARANTOR   < 0.5     to the left,  improve= 3.782059, (0 missing)
##       DURATION    < 11.5    to the right, improve= 3.766531, (0 missing)
## 
## Node number 8: 196 observations,    complexity param=0.01833333
##   predicted class=0  expected loss=0.377551  P(node) =0.196
##     class counts:   122    74
##    probabilities: 0.622 0.378 
##   left son=16 (36 obs) right son=17 (160 obs)
##   Primary splits:
##       DURATION         < 47.5    to the right, improve=5.023838, (0 missing)
##       USED_CAR         < 0.5     to the left,  improve=4.598639, (0 missing)
##       INSTALL_RATE     < 2.5     to the right, improve=2.682485, (0 missing)
##       AMOUNT           < 11788   to the right, improve=2.516732, (0 missing)
##       PRESENT_RESIDENT < 1.5     to the right, improve=1.984382, (0 missing)
##   Surrogate splits:
##       AMOUNT < 13319.5 to the right, agree=0.837, adj=0.111, (0 split)
## 
## Node number 9: 41 observations
##   predicted class=1  expected loss=0.2926829  P(node) =0.041
##     class counts:    12    29
##    probabilities: 0.293 0.707 
## 
## Node number 10: 28 observations
##   predicted class=0  expected loss=0.25  P(node) =0.028
##     class counts:    21     7
##    probabilities: 0.750 0.250 
## 
## Node number 11: 278 observations,    complexity param=0.014
##   predicted class=1  expected loss=0.3057554  P(node) =0.278
##     class counts:    85   193
##    probabilities: 0.306 0.694 
##   left son=22 (241 obs) right son=23 (37 obs)
##   Primary splits:
##       OBS.        < 120.5   to the right, improve=5.407923, (0 missing)
##       AMOUNT      < 7491.5  to the right, improve=4.366338, (0 missing)
##       DURATION    < 11.5    to the right, improve=3.840775, (0 missing)
##       REAL_ESTATE < 0.5     to the left,  improve=3.589042, (0 missing)
##       HISTORY     < 2.5     to the left,  improve=2.954088, (0 missing)
## 
## Node number 16: 36 observations
##   predicted class=0  expected loss=0.1388889  P(node) =0.036
##     class counts:    31     5
##    probabilities: 0.861 0.139
##
## Node number 17: 160 observations,    complexity param=0.01833333
##   predicted class=0  expected loss=0.43125  P(node) =0.16
##     class counts:    91    69
##    probabilities: 0.569 0.431
##   left son=34 (137 obs) right son=35 (23 obs)
##   Primary splits:
##       USED_CAR     < 0.5     to the left,  improve=5.092387, (0 missing)
##       AMOUNT       < 2313    to the left,  improve=3.402464, (0 missing)
##       INSTALL_RATE < 2.5     to the right, improve=2.374236, (0 missing)
##       NEW_CAR      < 0.5     to the right, improve=2.000321, (0 missing)
##       AGE          < 57.5    to the left,  improve=1.711184, (0 missing)
##   Surrogate splits:
##       OBS. < 982.5   to the left,  agree=0.863, adj=0.043, (0 split)
##       AGE  < 62      to the left,  agree=0.863, adj=0.043, (0 split)
##
## Node number 22: 241 observations,    complexity param=0.014
##   predicted class=1  expected loss=0.3443983  P(node) =0.241
##     class counts:    83   158
##    probabilities: 0.344 0.656
##   left son=44 (7 obs) right son=45 (234 obs)
##   Primary splits:
##       AMOUNT      < 7491.5  to the right, improve=3.790803, (0 missing)
##       OBS.        < 933.5   to the left,  improve=3.525911, (0 missing)
##       GUARANTOR   < 0.5     to the left,  improve=3.309626, (0 missing)
##       DURATION    < 11.5    to the right, improve=3.180698, (0 missing)
##       REAL_ESTATE < 0.5     to the left,  improve=2.854868, (0 missing)
##
## Node number 23: 37 observations
##   predicted class=1  expected loss=0.05405405  P(node) =0.037
##     class counts:     2    35
##    probabilities: 0.054 0.946
##
## Node number 34: 137 observations
##   predicted class=0  expected loss=0.379562  P(node) =0.137
##     class counts:    85    52
##    probabilities: 0.620 0.380
##
## Node number 35: 23 observations
##   predicted class=1  expected loss=0.2608696  P(node) =0.023
##     class counts:     6    17
##    probabilities: 0.261 0.739
##
## Node number 44: 7 observations
##   predicted class=0  expected loss=0.1428571  P(node) =0.007
##     class counts:     6     1
##    probabilities: 0.857 0.143
##
## Node number 45: 234 observations,    complexity param=0.014
##   predicted class=1  expected loss=0.3290598  P(node) =0.234
##     class counts:    77   157
##    probabilities: 0.329 0.671
##   left son=90 (200 obs) right son=91 (34 obs)
##   Primary splits:
##       DURATION  < 8.5     to the right, improve=3.555963, (0 missing)
##       OBS.      < 933.5   to the left,  improve=3.224786, (0 missing)
##       GUARANTOR < 0.5     to the left,  improve=2.899666, (0 missing)
##       AMOUNT    < 1541.5  to the left,  improve=2.886843, (0 missing)
##       EDUCATION < 0.5     to the right, improve=2.874786, (0 missing)
##   Surrogate splits:
##       AMOUNT < 527.5   to the right, agree=0.876, adj=0.147, (0 split)
##
## Node number 90: 200 observations,    complexity param=0.014
##   predicted class=1  expected loss=0.365  P(node) =0.2
##     class counts:    73   127
##    probabilities: 0.365 0.635
##   left son=180 (85 obs) right son=181 (115 obs)
##   Primary splits:
##       AMOUNT           < 1423    to the left,  improve=4.928926, (0 missing)
##       OBS.             < 933.5   to the left,  improve=3.101534, (0 missing)
##       GUARANTOR        < 0.5     to the left,  improve=2.315746, (0 missing)
##       PRESENT_RESIDENT < 3.5     to the left,  improve=2.108112, (0 missing)
##       MALE_SINGLE      < 0.5     to the left,  improve=1.895859, (0 missing)
##   Surrogate splits:
##       INSTALL_RATE    < 3.5     to the right, agree=0.655, adj=0.188, (0 split)
##       DURATION        < 12.5    to the left,  agree=0.640, adj=0.153, (0 split)
##       JOB             < 1.5     to the left,  agree=0.625, adj=0.118, (0 split)
##       MALE_MAR_or_WID < 0.5     to the right, agree=0.605, adj=0.071, (0 split)
##       AGE             < 45      to the right, agree=0.605, adj=0.071, (0 split)
##
## Node number 91: 34 observations
##   predicted class=1  expected loss=0.1176471  P(node) =0.034
##     class counts:     4    30
##    probabilities: 0.118 0.882
##
## Node number 180: 85 observations,    complexity param=0.014
##   predicted class=1  expected loss=0.4941176  P(node) =0.085
##     class counts:    42    43
##    probabilities: 0.494 0.506
##   left son=360 (48 obs) right son=361 (37 obs)
##   Primary splits:
##       REAL_ESTATE      < 0.5     to the left,  improve=6.566190, (0 missing)
##       NUM_CREDITS      < 1.5     to the left,  improve=5.195552, (0 missing)
##       GUARANTOR        < 0.5     to the left,  improve=4.715579, (0 missing)
##       PRESENT_RESIDENT < 3.5     to the left,  improve=4.440830, (0 missing)
##       AGE              < 37.5    to the left,  improve=2.973182, (0 missing)
##   Surrogate splits:
##       RADIO.TV        < 0.5     to the left,  agree=0.729, adj=0.378, (0 split)
##       GUARANTOR       < 0.5     to the left,  agree=0.659, adj=0.216, (0 split)
##       JOB             < 1.5     to the right, agree=0.659, adj=0.216, (0 split)
##       AMOUNT          < 632     to the right, agree=0.624, adj=0.135, (0 split)
##       MALE_MAR_or_WID < 0.5     to the left,  agree=0.624, adj=0.135, (0 split)
##
## Node number 181: 115 observations
##   predicted class=1  expected loss=0.2695652  P(node) =0.115
##     class counts:    31    84
##    probabilities: 0.270 0.730
```

```
##
## Node number 360: 48 observations,    complexity param=0.01333333
##   predicted class=0  expected loss=0.3333333  P(node) =0.048
##     class counts:    32    16
##    probabilities: 0.667 0.333
##   left son=720 (34 obs) right son=721 (14 obs)
##   Primary splits:
##       AGE              < 37.5    to the left,  improve=3.787115, (0 missing)
##       NUM_CREDITS      < 1.5     to the left,  improve=3.555556, (0 missing)
##       PRESENT_RESIDENT < 2.5     to the left,  improve=2.711485, (0 missing)
##       AMOUNT           < 967     to the left,  improve=1.864802, (0 missing)
##       EMPLOYMENT       < 1.5     to the left,  improve=1.434174, (0 missing)
##   Surrogate splits:
##       EDUCATION        < 0.5     to the left,  agree=0.750, adj=0.143, (0 split)
##       NUM_CREDITS      < 1.5     to the left,  agree=0.750, adj=0.143, (0 split)
##       JOB              < 0.5     to the right, agree=0.750, adj=0.143, (0 split)
##       NUM_DEPENDENTS   < 1.5     to the left,  agree=0.750, adj=0.143, (0 split)
##       PRESENT_RESIDENT < 3.5     to the left,  agree=0.729, adj=0.071, (0 split)
##
## Node number 361: 37 observations
##   predicted class=1  expected loss=0.2702703  P(node) =0.037
##     class counts:    10    27
##    probabilities: 0.270 0.730
##
## Node number 720: 34 observations
##   predicted class=0  expected loss=0.2058824  P(node) =0.034
##     class counts:    27     7
##    probabilities: 0.794 0.206
##
## Node number 721: 14 observations
##   predicted class=1  expected loss=0.3571429  P(node) =0.014
##     class counts:     5     9
##    probabilities: 0.357 0.643
```

The model is not working yet.

```
par(pty = "s", mar = c(1, 1, 1, 1))
plot(german.ct, cex = 1)
text(german.ct, cex = 0.6)
```



Fitting another model :

```
# Logistic regression to see the significant variables (not working)
mod1 <- glm(RESPONSE~., data = German_credit, family= binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = RESPONSE ~ ., family = binomial, data = German_credit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6301  -0.7228   0.3889   0.7030   2.3722
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.057e+00  8.704e-01   1.214  0.22479
## OBS.               -1.623e-04  2.884e-04  -0.563  0.57356
## CHK_ACCT            5.633e-01  7.254e-02   7.766 8.12e-15 ***
## DURATION           -2.719e-02  9.046e-03  -3.005  0.00265 **
## HISTORY             3.968e-01  8.980e-02   4.419 9.93e-06 ***
## NEW_CAR            -7.996e-01  3.811e-01  -2.098  0.03588 *
## USED_CAR            8.218e-01  4.782e-01   1.718  0.08571 .
## FURNITURE          -4.223e-02  3.952e-01  -0.107  0.91490
## RADIO.TV            6.264e-02  3.846e-01   0.163  0.87061
## EDUCATION          -9.249e-01  4.952e-01  -1.868  0.06182 .
## RETRAINING         -8.732e-02  4.376e-01  -0.200  0.84182
## AMOUNT             -1.168e-04  4.272e-05  -2.734  0.00625 **
## SAV_ACCT          2.497e-01  6.066e-02   4.116 3.85e-05 ***
## EMPLOYMENT        1.168e-01  7.469e-02   1.564  0.11781    
## INSTALL_RATE     -3.171e-01  8.659e-02  -3.662  0.00025 ***
## MALE_DIV         -3.443e-01  3.814e-01  -0.903  0.36663    
## MALE_SINGLE       5.378e-01  2.051e-01   2.622  0.00874 ** 
## MALE_MAR_or_WID   1.069e-01  3.046e-01   0.351  0.72572    
## CO.APPLICANT     -3.494e-01  3.989e-01  -0.876  0.38113    
## GUARANTOR         9.451e-01  4.146e-01   2.279  0.02265 *  
## PRESENT_RESIDENT -1.242e-02  8.403e-02  -0.148  0.88247    
## REAL_ESTATE       1.997e-01  2.096e-01   0.953  0.34083    
## PROP_UNKN_NONE   -5.569e-01  3.735e-01  -1.491  0.13595    
## AGE               1.211e-02  8.383e-03   1.445  0.14843    
## OTHER_INSTALL    -6.310e-01  2.045e-01  -3.085  0.00203 ** 
## RENT             -6.277e-01  4.608e-01  -1.362  0.17313    
## OWN_RES          -2.222e-01  4.360e-01  -0.510  0.61030    
## NUM_CREDITS      -2.238e-01  1.663e-01  -1.346  0.17833    
## JOB              -3.325e-02  1.427e-01  -0.233  0.81569    
## NUM_DEPENDENTS   -2.480e-01  2.461e-01  -1.008  0.31349    
## TELEPHONE         3.507e-01  1.955e-01   1.794  0.07288 .  
## FOREIGN           1.458e+00  6.243e-01   2.336  0.01951 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1221.73  on 999  degrees of freedom
## Residual deviance:  907.75  on 968  degrees of freedom
## AIC: 971.75
## 
## Number of Fisher Scoring iterations: 5
```