data mining

Suneeth Ravilla

4/15/2020

setwd("C:/Users/rvssu/Downloads/DataMining")  
source("myfunctions.R")  
source("confusionMatrixFunctions.R")

GERMAN\_DATA = read.csv("C:/Users/rvssu/Downloads/DataMining/Project/GERMAN\_DATA.csv")

GERMAN\_DATA[!complete.cases(GERMAN\_DATA),]

## [1] CHK\_ACCT\_ST DUR CRED\_HIST   
## [4] PURPOSE CRED\_AMT SAV\_ACCT\_BOND   
## [7] EMPLYMT\_ST INST\_RT\_PER\_DISP\_INCM PERS\_ST\_SEX   
## [10] COAPP\_GURNTR DUR\_RES PROPERTY   
## [13] AGE OTHR\_INSTL HOUS\_ST   
## [16] NUM\_CRED JOB NUM\_PEOP\_LIABL   
## [19] PHONE FRGN\_WORKR Y   
## <0 rows> (or 0-length row.names)

summary(GERMAN\_DATA)

## CHK\_ACCT\_ST DUR CRED\_HIST PURPOSE CRED\_AMT   
## A11:274 Min. : 4.0 A30: 40 A43 :280 Min. : 250   
## A12:269 1st Qu.:12.0 A31: 49 A40 :234 1st Qu.: 1366   
## A13: 63 Median :18.0 A32:530 A42 :181 Median : 2320   
## A14:394 Mean :20.9 A33: 88 A41 :103 Mean : 3271   
## 3rd Qu.:24.0 A34:293 A49 : 97 3rd Qu.: 3972   
## Max. :72.0 A46 : 50 Max. :18424   
## (Other): 55   
## SAV\_ACCT\_BOND EMPLYMT\_ST INST\_RT\_PER\_DISP\_INCM PERS\_ST\_SEX COAPP\_GURNTR  
## A61:603 A71: 62 Min. :1.000 A91: 50 A101:907   
## A62:103 A72:172 1st Qu.:2.000 A92:310 A102: 41   
## A63: 63 A73:339 Median :3.000 A93:548 A103: 52   
## A64: 48 A74:174 Mean :2.973 A94: 92   
## A65:183 A75:253 3rd Qu.:4.000   
## Max. :4.000   
##   
## DUR\_RES PROPERTY AGE OTHR\_INSTL HOUS\_ST   
## Min. :1.000 A121:282 Min. :19.00 A141:139 A151:179   
## 1st Qu.:2.000 A122:232 1st Qu.:27.00 A142: 47 A152:713   
## Median :3.000 A123:332 Median :33.00 A143:814 A153:108   
## Mean :2.845 A124:154 Mean :35.55   
## 3rd Qu.:4.000 3rd Qu.:42.00   
## Max. :4.000 Max. :75.00   
##   
## NUM\_CRED JOB NUM\_PEOP\_LIABL PHONE FRGN\_WORKR Y   
## Min. :1.000 A171: 22 Min. :1.000 A191:596 A201:963 Min. :1.0   
## 1st Qu.:1.000 A172:200 1st Qu.:1.000 A192:404 A202: 37 1st Qu.:1.0   
## Median :1.000 A173:630 Median :1.000 Median :1.0   
## Mean :1.407 A174:148 Mean :1.155 Mean :1.3   
## 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:2.0   
## Max. :4.000 Max. :2.000 Max. :2.0   
##

# Logistic model

data3 = GERMAN\_DATA

data3$PHONE = factor(data3$PHONE, levels = c('A191','A192'), labels = c(1,2))  
data3$FRGN\_WORKR = factor(data3$FRGN\_WORKR, levels = c('A201','A202'), labels = c(1,2))  
  
data3$COAPP\_GURNTR = factor(data3$COAPP\_GURNTR, levels = c('A101','A102','A103'), labels = c(1,2,3))  
data3$OTHR\_INSTL = factor(data3$OTHR\_INSTL, levels = c('A141','A142','A143'), labels = c(1,2,3))  
data3$HOUS\_ST = factor(data3$HOUS\_ST, levels = c('A151','A152','A153'), labels = c(1,2,3))  
  
data3$CHK\_ACCT\_ST = factor(data3$CHK\_ACCT\_ST, levels = c('A11','A12','A13','A14'), labels = c(1,2,3,4))  
data3$PERS\_ST\_SEX = factor(data3$PERS\_ST\_SEX, levels = c('A91','A92','A93','A94'), labels = c(1,2,3,4))  
data3$PROPERTY = factor(data3$PROPERTY, levels = c('A121','A122','A123','A124'), labels = c(1,2,3,4))  
data3$JOB = factor(data3$JOB, levels = c('A171','A172','A173','A174'), labels = c(1,2,3,4))  
  
data3$CRED\_HIST = factor(data3$CRED\_HIST, levels = c('A30','A31','A32','A33','A34'), labels = c(1,2,3,4,5))  
data3$SAV\_ACCT\_BOND = factor(data3$SAV\_ACCT\_BOND, levels = c('A61','A62','A63','A64','A65'), labels = c(1,2,3,4,5))  
data3$EMPLYMT\_ST = factor(data3$EMPLYMT\_ST, levels = c('A71','A72','A73','A74','A75'), labels = c(1,2,3,4,5))  
  
data3$PURPOSE = factor(data3$PURPOSE, levels = c('A40','A41','A42','A43','A44','A45','A46','A48','A49','A410'), labels = c(1,2,3,4,5,6,7,8,9,10))

data3$Y = ifelse(data3$Y =="2",1,0)  
data3$Y = as.factor(data3$Y)

set.seed(0)  
p2 <- partition.2(data3, 0.7)  
data3.train <- p2$data.train  
data3.val <- p2$data.val

logistic = glm(Y~.-PHONE - PERS\_ST\_SEX,family=binomial(link='logit'),data = data3.train)  
summary(logistic)

##   
## Call:  
## glm(formula = Y ~ . - PHONE - PERS\_ST\_SEX, family = binomial(link = "logit"),   
## data = data3.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1904 -0.7304 -0.3977 0.7097 2.4682   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.029e+00 1.330e+00 -0.774 0.43908   
## CHK\_ACCT\_ST2 -4.800e-01 2.674e-01 -1.795 0.07262 .   
## CHK\_ACCT\_ST3 -9.472e-01 4.423e-01 -2.141 0.03224 \*   
## CHK\_ACCT\_ST4 -1.541e+00 2.713e-01 -5.680 1.35e-08 \*\*\*  
## DUR 2.978e-02 1.167e-02 2.551 0.01074 \*   
## CRED\_HIST2 4.397e-01 6.526e-01 0.674 0.50043   
## CRED\_HIST3 -4.746e-01 5.147e-01 -0.922 0.35647   
## CRED\_HIST4 -4.052e-01 5.537e-01 -0.732 0.46424   
## CRED\_HIST5 -1.012e+00 5.249e-01 -1.928 0.05388 .   
## PURPOSE2 -1.355e+00 4.289e-01 -3.159 0.00158 \*\*   
## PURPOSE3 -6.718e-01 3.092e-01 -2.173 0.02979 \*   
## PURPOSE4 -8.807e-01 2.879e-01 -3.060 0.00222 \*\*   
## PURPOSE5 -3.101e-01 9.917e-01 -0.313 0.75451   
## PURPOSE6 -2.400e-01 6.617e-01 -0.363 0.71685   
## PURPOSE7 -3.096e-01 4.788e-01 -0.647 0.51790   
## PURPOSE8 -1.537e+01 4.954e+02 -0.031 0.97524   
## PURPOSE9 -7.097e-01 4.031e-01 -1.761 0.07828 .   
## PURPOSE10 -2.903e+00 1.370e+00 -2.119 0.03410 \*   
## CRED\_AMT 8.842e-05 5.762e-05 1.534 0.12493   
## SAV\_ACCT\_BOND2 -1.238e-01 3.339e-01 -0.371 0.71083   
## SAV\_ACCT\_BOND3 1.499e-01 4.277e-01 0.350 0.72602   
## SAV\_ACCT\_BOND4 -1.407e+00 7.833e-01 -1.797 0.07239 .   
## SAV\_ACCT\_BOND5 -9.085e-01 3.283e-01 -2.767 0.00566 \*\*   
## EMPLYMT\_ST2 6.392e-02 5.060e-01 0.126 0.89947   
## EMPLYMT\_ST3 -2.024e-01 4.840e-01 -0.418 0.67590   
## EMPLYMT\_ST4 -9.726e-01 5.375e-01 -1.810 0.07034 .   
## EMPLYMT\_ST5 -3.016e-01 4.923e-01 -0.613 0.54011   
## INST\_RT\_PER\_DISP\_INCM 3.208e-01 1.068e-01 3.003 0.00267 \*\*   
## COAPP\_GURNTR2 3.390e-01 4.984e-01 0.680 0.49642   
## COAPP\_GURNTR3 -7.037e-01 5.215e-01 -1.349 0.17718   
## DUR\_RES -5.262e-02 1.036e-01 -0.508 0.61151   
## PROPERTY2 1.547e-01 3.047e-01 0.508 0.61160   
## PROPERTY3 2.338e-01 2.805e-01 0.833 0.40458   
## PROPERTY4 6.083e-01 5.436e-01 1.119 0.26313   
## AGE -1.739e-02 1.127e-02 -1.542 0.12301   
## OTHR\_INSTL2 -1.232e-01 4.861e-01 -0.253 0.79999   
## OTHR\_INSTL3 -5.500e-01 2.916e-01 -1.886 0.05928 .   
## HOUS\_ST2 -5.374e-01 2.750e-01 -1.954 0.05071 .   
## HOUS\_ST3 -5.659e-01 5.957e-01 -0.950 0.34214   
## NUM\_CRED 2.482e-01 2.167e-01 1.146 0.25199   
## JOB2 1.612e+00 9.636e-01 1.673 0.09438 .   
## JOB3 1.529e+00 9.366e-01 1.632 0.10266   
## JOB4 1.140e+00 9.595e-01 1.188 0.23468   
## NUM\_PEOP\_LIABL 1.095e-01 2.803e-01 0.391 0.69607   
## FRGN\_WORKR2 -2.445e+00 1.091e+00 -2.242 0.02496 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 832.02 on 699 degrees of freedom  
## Residual deviance: 633.62 on 655 degrees of freedom  
## AIC: 723.62  
##   
## Number of Fisher Scoring iterations: 14

library(stats)  
logistic.m2.step = step(logistic)

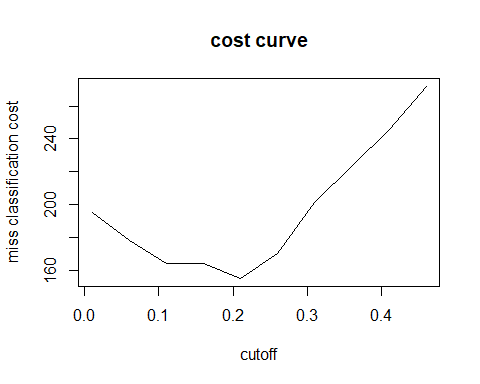
## Start: AIC=723.62  
## Y ~ (CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + PERS\_ST\_SEX + COAPP\_GURNTR +   
## DUR\_RES + PROPERTY + AGE + OTHR\_INSTL + HOUS\_ST + NUM\_CRED +   
## JOB + NUM\_PEOP\_LIABL + PHONE + FRGN\_WORKR) - PHONE - PERS\_ST\_SEX  
##   
## Df Deviance AIC  
## - PROPERTY 3 635.11 719.11  
## - NUM\_PEOP\_LIABL 1 633.77 721.77  
## - DUR\_RES 1 633.88 721.88  
## - JOB 3 637.94 721.94  
## - COAPP\_GURNTR 2 636.20 722.20  
## - NUM\_CRED 1 634.93 722.93  
## - HOUS\_ST 2 637.53 723.53  
## - OTHR\_INSTL 2 637.59 723.59  
## <none> 633.62 723.62  
## - CRED\_AMT 1 635.99 723.99  
## - AGE 1 636.06 724.06  
## - EMPLYMT\_ST 4 642.89 724.89  
## - CRED\_HIST 4 643.64 725.64  
## - SAV\_ACCT\_BOND 4 646.16 728.16  
## - DUR 1 640.17 728.17  
## - PURPOSE 9 657.94 729.94  
## - FRGN\_WORKR 1 642.32 730.32  
## - INST\_RT\_PER\_DISP\_INCM 1 642.94 730.94  
## - CHK\_ACCT\_ST 3 670.34 754.34  
##   
## Step: AIC=719.11  
## Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + COAPP\_GURNTR + DUR\_RES +   
## AGE + OTHR\_INSTL + HOUS\_ST + NUM\_CRED + JOB + NUM\_PEOP\_LIABL +   
## FRGN\_WORKR  
##   
## Df Deviance AIC  
## - JOB 3 639.03 717.03  
## - NUM\_PEOP\_LIABL 1 635.23 717.23  
## - DUR\_RES 1 635.32 717.32  
## - COAPP\_GURNTR 2 638.19 718.19  
## - NUM\_CRED 1 636.32 718.32  
## <none> 635.11 719.11  
## - HOUS\_ST 2 639.40 719.40  
## - AGE 1 637.58 719.58  
## - OTHR\_INSTL 2 639.64 719.64  
## - CRED\_AMT 1 638.05 720.05  
## - EMPLYMT\_ST 4 644.25 720.25  
## - CRED\_HIST 4 645.22 721.22  
## - DUR 1 642.11 724.11  
## - SAV\_ACCT\_BOND 4 648.12 724.12  
## - FRGN\_WORKR 1 643.61 725.61  
## - PURPOSE 9 659.99 725.99  
## - INST\_RT\_PER\_DISP\_INCM 1 645.19 727.19  
## - CHK\_ACCT\_ST 3 673.15 751.15  
##   
## Step: AIC=717.03  
## Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + COAPP\_GURNTR + DUR\_RES +   
## AGE + OTHR\_INSTL + HOUS\_ST + NUM\_CRED + NUM\_PEOP\_LIABL +   
## FRGN\_WORKR  
##   
## Df Deviance AIC  
## - DUR\_RES 1 639.10 715.10  
## - NUM\_PEOP\_LIABL 1 639.20 715.20  
## - COAPP\_GURNTR 2 641.65 715.65  
## - NUM\_CRED 1 639.93 715.93  
## - HOUS\_ST 2 642.72 716.72  
## <none> 639.03 717.03  
## - OTHR\_INSTL 2 643.43 717.43  
## - EMPLYMT\_ST 4 647.60 717.60  
## - CRED\_AMT 1 641.71 717.71  
## - AGE 1 641.73 717.73  
## - CRED\_HIST 4 648.62 718.62  
## - DUR 1 646.00 722.00  
## - SAV\_ACCT\_BOND 4 652.50 722.50  
## - FRGN\_WORKR 1 647.18 723.18  
## - PURPOSE 9 664.72 724.72  
## - INST\_RT\_PER\_DISP\_INCM 1 648.73 724.73  
## - CHK\_ACCT\_ST 3 677.46 749.46  
##   
## Step: AIC=715.1  
## Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + COAPP\_GURNTR + AGE +   
## OTHR\_INSTL + HOUS\_ST + NUM\_CRED + NUM\_PEOP\_LIABL + FRGN\_WORKR  
##   
## Df Deviance AIC  
## - NUM\_PEOP\_LIABL 1 639.25 713.25  
## - COAPP\_GURNTR 2 641.69 713.69  
## - NUM\_CRED 1 639.97 713.97  
## - HOUS\_ST 2 642.87 714.87  
## <none> 639.10 715.10  
## - OTHR\_INSTL 2 643.48 715.48  
## - CRED\_AMT 1 641.86 715.86  
## - EMPLYMT\_ST 4 647.90 715.90  
## - AGE 1 641.91 715.91  
## - CRED\_HIST 4 648.77 716.77  
## - DUR 1 646.02 720.02  
## - SAV\_ACCT\_BOND 4 652.67 720.67  
## - FRGN\_WORKR 1 647.19 721.19  
## - INST\_RT\_PER\_DISP\_INCM 1 648.73 722.73  
## - PURPOSE 9 664.84 722.84  
## - CHK\_ACCT\_ST 3 677.47 747.47  
##   
## Step: AIC=713.25  
## Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + COAPP\_GURNTR + AGE +   
## OTHR\_INSTL + HOUS\_ST + NUM\_CRED + FRGN\_WORKR  
##   
## Df Deviance AIC  
## - COAPP\_GURNTR 2 641.79 711.79  
## - NUM\_CRED 1 640.18 712.18  
## - HOUS\_ST 2 643.01 713.01  
## <none> 639.25 713.25  
## - OTHR\_INSTL 2 643.70 713.70  
## - CRED\_AMT 1 641.96 713.96  
## - EMPLYMT\_ST 4 647.97 713.97  
## - AGE 1 641.98 713.98  
## - CRED\_HIST 4 649.02 715.02  
## - DUR 1 646.17 718.17  
## - SAV\_ACCT\_BOND 4 652.85 718.85  
## - FRGN\_WORKR 1 647.25 719.25  
## - INST\_RT\_PER\_DISP\_INCM 1 648.75 720.75  
## - PURPOSE 9 665.03 721.03  
## - CHK\_ACCT\_ST 3 678.26 746.26  
##   
## Step: AIC=711.79  
## Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + AGE + OTHR\_INSTL + HOUS\_ST +   
## NUM\_CRED + FRGN\_WORKR  
##   
## Df Deviance AIC  
## - NUM\_CRED 1 642.72 710.72  
## - HOUS\_ST 2 645.67 711.67  
## - OTHR\_INSTL 2 645.71 711.71  
## <none> 641.79 711.79  
## - CRED\_AMT 1 644.67 712.67  
## - AGE 1 644.71 712.71  
## - EMPLYMT\_ST 4 651.24 713.24  
## - CRED\_HIST 4 651.28 713.28  
## - DUR 1 648.85 716.85  
## - SAV\_ACCT\_BOND 4 655.52 717.52  
## - FRGN\_WORKR 1 650.95 718.95  
## - INST\_RT\_PER\_DISP\_INCM 1 651.82 719.82  
## - PURPOSE 9 668.71 720.71  
## - CHK\_ACCT\_ST 3 680.24 744.24  
##   
## Step: AIC=710.72  
## Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT + SAV\_ACCT\_BOND +   
## EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + AGE + OTHR\_INSTL + HOUS\_ST +   
## FRGN\_WORKR  
##   
## Df Deviance AIC  
## <none> 642.72 710.72  
## - HOUS\_ST 2 646.75 710.75  
## - OTHR\_INSTL 2 646.93 710.93  
## - CRED\_HIST 4 651.30 711.30  
## - AGE 1 645.51 711.51  
## - CRED\_AMT 1 645.85 711.85  
## - EMPLYMT\_ST 4 651.91 711.91  
## - DUR 1 649.37 715.37  
## - SAV\_ACCT\_BOND 4 656.49 716.49  
## - FRGN\_WORKR 1 652.12 718.12  
## - INST\_RT\_PER\_DISP\_INCM 1 652.78 718.78  
## - PURPOSE 9 669.85 719.85  
## - CHK\_ACCT\_ST 3 681.03 743.03

summary(logistic.m2.step)

##   
## Call:  
## glm(formula = Y ~ CHK\_ACCT\_ST + DUR + CRED\_HIST + PURPOSE + CRED\_AMT +   
## SAV\_ACCT\_BOND + EMPLYMT\_ST + INST\_RT\_PER\_DISP\_INCM + AGE +   
## OTHR\_INSTL + HOUS\_ST + FRGN\_WORKR, family = binomial(link = "logit"),   
## data = data3.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0021 -0.7284 -0.4047 0.7498 2.5435   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.135e-01 8.755e-01 0.701 0.483449   
## CHK\_ACCT\_ST2 -5.679e-01 2.580e-01 -2.201 0.027761 \*   
## CHK\_ACCT\_ST3 -1.048e+00 4.335e-01 -2.417 0.015629 \*   
## CHK\_ACCT\_ST4 -1.553e+00 2.663e-01 -5.832 5.47e-09 \*\*\*  
## DUR 2.876e-02 1.118e-02 2.573 0.010085 \*   
## CRED\_HIST2 2.400e-01 6.198e-01 0.387 0.698578   
## CRED\_HIST3 -5.791e-01 4.893e-01 -1.184 0.236559   
## CRED\_HIST4 -3.247e-01 5.436e-01 -0.597 0.550307   
## CRED\_HIST5 -9.542e-01 5.154e-01 -1.851 0.064118 .   
## PURPOSE2 -1.404e+00 4.185e-01 -3.354 0.000796 \*\*\*  
## PURPOSE3 -6.257e-01 2.991e-01 -2.092 0.036445 \*   
## PURPOSE4 -8.976e-01 2.833e-01 -3.168 0.001532 \*\*   
## PURPOSE5 -4.243e-01 9.972e-01 -0.426 0.670457   
## PURPOSE6 -1.477e-01 6.445e-01 -0.229 0.818756   
## PURPOSE7 -2.282e-01 4.708e-01 -0.485 0.627862   
## PURPOSE8 -1.513e+01 4.972e+02 -0.030 0.975729   
## PURPOSE9 -6.692e-01 3.919e-01 -1.708 0.087719 .   
## PURPOSE10 -3.108e+00 1.324e+00 -2.348 0.018865 \*   
## CRED\_AMT 9.491e-05 5.373e-05 1.766 0.077355 .   
## SAV\_ACCT\_BOND2 6.128e-02 3.197e-01 0.192 0.848004   
## SAV\_ACCT\_BOND3 1.295e-01 4.195e-01 0.309 0.757515   
## SAV\_ACCT\_BOND4 -1.376e+00 7.766e-01 -1.772 0.076366 .   
## SAV\_ACCT\_BOND5 -9.312e-01 3.239e-01 -2.875 0.004045 \*\*   
## EMPLYMT\_ST2 3.838e-01 4.618e-01 0.831 0.405962   
## EMPLYMT\_ST3 1.280e-01 4.321e-01 0.296 0.766975   
## EMPLYMT\_ST4 -6.441e-01 4.940e-01 -1.304 0.192271   
## EMPLYMT\_ST5 4.027e-02 4.498e-01 0.090 0.928659   
## INST\_RT\_PER\_DISP\_INCM 3.206e-01 1.030e-01 3.114 0.001848 \*\*   
## AGE -1.800e-02 1.092e-02 -1.649 0.099112 .   
## OTHR\_INSTL2 -5.662e-02 4.825e-01 -0.117 0.906575   
## OTHR\_INSTL3 -5.365e-01 2.835e-01 -1.893 0.058413 .   
## HOUS\_ST2 -4.768e-01 2.553e-01 -1.868 0.061805 .   
## HOUS\_ST3 -1.284e-01 3.918e-01 -0.328 0.743151   
## FRGN\_WORKR2 -2.477e+00 1.080e+00 -2.293 0.021875 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 832.02 on 699 degrees of freedom  
## Residual deviance: 642.72 on 666 degrees of freedom  
## AIC: 710.72  
##   
## Number of Fisher Scoring iterations: 14

pred.val <- predict(logistic.m2.step, newdata = data3.val,type = "response")

# Asymmetric missclassification  
cost.ratio <- 5  
# create a vector for cutoff   
cutoff <- seq(0.01, 0.5, 0.05)  
# create an empty vector of same lengthmiss  
missclassification.cost <- rep(NA, length(cutoff))  
# calculate missclassification cost for each cutoff value  
for(i in 1:length(cutoff)){  
 res.conf <- conf(data3.val$Y,  
pred.val, cutoff[i])  
missclassification.cost[i] <- res.conf$conf.mat[2,1] +res.conf$conf.mat[1,2]\*cost.ratio  
}  
# create a plot for cutoff vs. cost  
plot(cutoff, missclassification.cost, main = "cost curve",xlab = "cutoff", ylab = "miss classification cost",type = "l")



conf(data3.val$Y,pred.val,0.21)

## $conf.mat  
## actual  
## pred 0 1  
## 0 142 20  
## 1 55 83  
##   
## $err  
## [1] 0.25  
##   
## $accuracy  
## [1] 0.75  
##   
## $sensitivity  
## [1] 0.8058252  
##   
## $specificity  
## [1] 0.7208122

library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

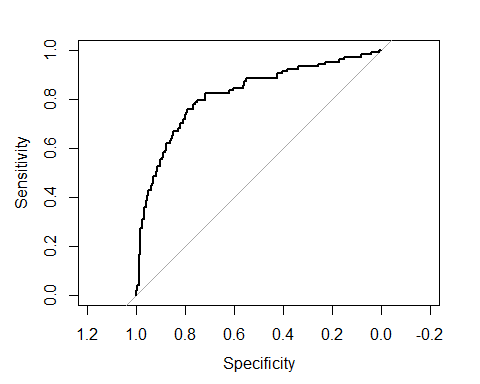
## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

r= roc(data3.val$Y, pred.val)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot.roc(r)



auc(r)

## Area under the curve: 0.8197

AUC value: 0.8132

# Random Forest

data4=GERMAN\_DATA

data4$Y = ifelse(data4$Y =="2",1,0)  
data4$Y = as.factor(data4$Y)

data4$PHONE = factor(data4$PHONE, levels = c('A191','A192'), labels = c(1,2))  
data4$FRGN\_WORKR = factor(data4$FRGN\_WORKR, levels = c('A201','A202'), labels = c(1,2))  
  
data4$COAPP\_GURNTR = factor(data4$COAPP\_GURNTR, levels = c('A101','A102','A103'), labels = c(1,2,3))  
data4$OTHR\_INSTL = factor(data4$OTHR\_INSTL, levels = c('A141','A142','A143'), labels = c(1,2,3))  
data4$HOUS\_ST = factor(data4$HOUS\_ST, levels = c('A151','A152','A153'), labels = c(1,2,3))  
  
data4$CHK\_ACCT\_ST = factor(data4$CHK\_ACCT\_ST, levels = c('A11','A12','A13','A14'), labels = c(1,2,3,4))  
data4$PERS\_ST\_SEX = factor(data4$PERS\_ST\_SEX, levels = c('A91','A92','A93','A94'), labels = c(1,2,3,4))  
data4$PROPERTY = factor(data4$PROPERTY, levels = c('A121','A122','A123','A124'), labels = c(1,2,3,4))  
data4$JOB = factor(data4$JOB, levels = c('A171','A172','A173','A174'), labels = c(1,2,3,4))  
  
data4$CRED\_HIST = factor(data4$CRED\_HIST, levels = c('A30','A31','A32','A33','A34'), labels = c(1,2,3,4,5))  
data4$SAV\_ACCT\_BOND = factor(data4$SAV\_ACCT\_BOND, levels = c('A61','A62','A63','A64','A65'), labels = c(1,2,3,4,5))  
data4$EMPLYMT\_ST = factor(data4$EMPLYMT\_ST, levels = c('A71','A72','A73','A74','A75'), labels = c(1,2,3,4,5))  
  
data4$PURPOSE = factor(data4$PURPOSE, levels = c('A40','A41','A42','A43','A44','A45','A46','A48','A49','A410'), labels = c(1,2,3,4,5,6,7,8,9,10))

library(scales)  
data4$DUR = rescale(data4$DUR)  
data4$INST\_RT\_PER\_DISP\_INCM = rescale(data4$INST\_RT\_PER\_DISP\_INCM)  
data4$DUR\_RES = rescale(data4$DUR\_RES)  
data4$AGE = rescale(data4$AGE)  
data4$NUM\_CRED= rescale(data4$NUM\_CRED)  
data4$NUM\_PEOP\_LIABL = rescale(data4$NUM\_PEOP\_LIABL)  
data4$CRED\_AMT = rescale(data4$CRED\_AMT)

set.seed(0)  
p2 <- partition.2(data4, 0.7)  
data4.train <- p2$data.train  
data4.val <- p2$data.val

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

set.seed(123)  
rndfrt= randomForest(Y~.-PHONE -PERS\_ST\_SEX,data=data4.train,ntree=500,type = "classification")  
rndfrt

##   
## Call:  
## randomForest(formula = Y ~ . - PHONE - PERS\_ST\_SEX, data = data4.train, ntree = 500, type = "classification")   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 24%  
## Confusion matrix:  
## 0 1 class.error  
## 0 470 33 0.06560636  
## 1 135 62 0.68527919

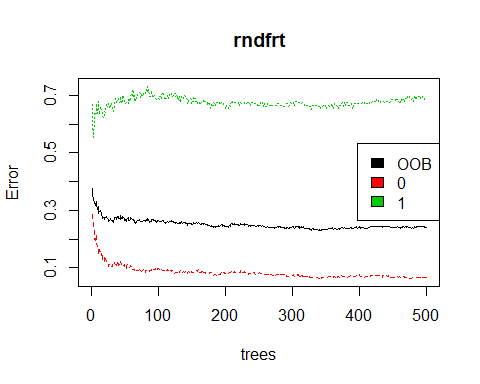
err\_rate = rndfrt$err.rate  
head(err\_rate)

## OOB 0 1  
## [1,] 0.3764706 0.2871795 0.6666667  
## [2,] 0.3574879 0.2682119 0.5982143  
## [3,] 0.3333333 0.2519894 0.5539568  
## [4,] 0.3310463 0.2203791 0.6211180  
## [5,] 0.3279743 0.2101770 0.6411765  
## [6,] 0.3176292 0.1970650 0.6353591

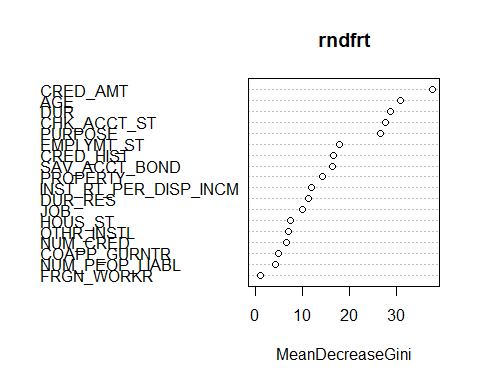
oob\_err <- err\_rate[nrow(err\_rate), "OOB"]  
print(oob\_err)

## OOB   
## 0.24

plot(rndfrt)  
legend(x = "right",   
 legend = colnames(err\_rate),  
 fill = 1:ncol(err\_rate))



varImpPlot(rndfrt)



rndfrt\_pred = predict(rndfrt, newdata = data4.val,type = "response")

cm = conf(rndfrt\_pred,data4.val$Y)  
print(cm)

## $conf.mat  
## actual  
## pred 0 1  
## 0 189 8  
## 1 68 35  
##   
## $err  
## [1] 0.2533333  
##   
## $accuracy  
## [1] 0.7466667  
##   
## $sensitivity  
## [1] 0.8139535  
##   
## $specificity  
## [1] 0.7354086

paste0("Test Accuracy: ", cm$accuracy)

## [1] "Test Accuracy: 0.746666666666667"

paste0("OOB Accuracy: ", 1 - oob\_err)

## [1] "OOB Accuracy: 0.76"

rndfrt\_predprob = predict(rndfrt,data4.val,type = "prob",probability =TRUE)  
auc(ifelse(data4.val$Y == "0",1, 0), rndfrt\_predprob[,"1"])

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

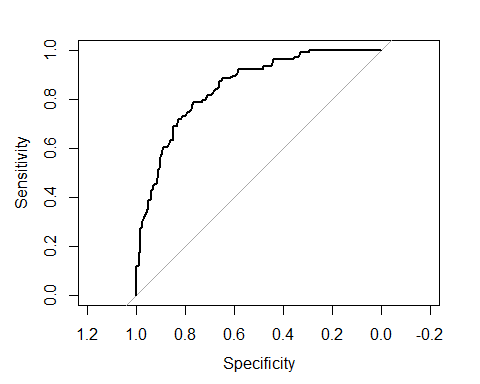
## Area under the curve: 0.8496

result.roc <- roc(data4.val$Y, rndfrt\_predprob[,"1"]) # Draw ROC curve.

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot(result.roc)



# Decision tree

data5 = GERMAN\_DATA

data5$Y = ifelse(data5$Y =="2",1,0)  
data5$Y = as.factor(data5$Y)

data5 = data5[,-c(9,19)]

data5$FRGN\_WORKR = factor(data5$FRGN\_WORKR, levels = c('A201','A202'), labels = c(1,2))  
  
data5$COAPP\_GURNTR = factor(data5$COAPP\_GURNTR, levels = c('A101','A102','A103'), labels = c(1,2,3))  
data5$OTHR\_INSTL = factor(data5$OTHR\_INSTL, levels = c('A141','A142','A143'), labels = c(1,2,3))  
data5$HOUS\_ST = factor(data5$HOUS\_ST, levels = c('A151','A152','A153'), labels = c(1,2,3))  
  
data5$CHK\_ACCT\_ST = factor(data5$CHK\_ACCT\_ST, levels = c('A11','A12','A13','A14'), labels = c(1,2,3,4))  
data5$PROPERTY = factor(data5$PROPERTY, levels = c('A121','A122','A123','A124'), labels = c(1,2,3,4))  
data5$JOB = factor(data5$JOB, levels = c('A171','A172','A173','A174'), labels = c(1,2,3,4))  
  
data5$CRED\_HIST = factor(data5$CRED\_HIST, levels = c('A30','A31','A32','A33','A34'), labels = c(1,2,3,4,5))  
data5$SAV\_ACCT\_BOND = factor(data5$SAV\_ACCT\_BOND, levels = c('A61','A62','A63','A64','A65'), labels = c(1,2,3,4,5))  
data5$EMPLYMT\_ST = factor(data5$EMPLYMT\_ST, levels = c('A71','A72','A73','A74','A75'), labels = c(1,2,3,4,5))  
  
data5$PURPOSE = factor(data5$PURPOSE, levels = c('A40','A41','A42','A43','A44','A45','A46','A48','A49','A410'), labels = c(1,2,3,4,5,6,7,8,9,10))

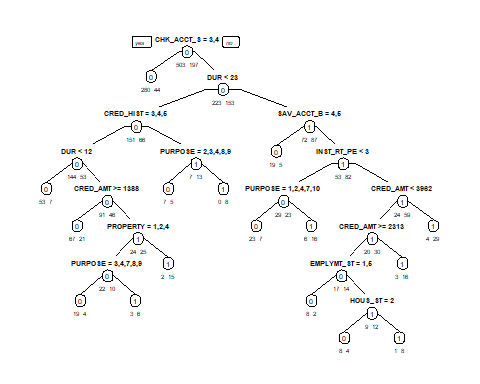
set.seed(0)  
p2 <- partition.2(data5, 0.7)  
data5.train <- p2$data.train  
data5.val <- p2$data.val

library(rpart)  
library(rpart.plot)

ct1 <- rpart(Y ~ . , data = data5.train, method = "class")  
ct1

## n= 700   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 700 197 0 (0.7185714 0.2814286)   
## 2) CHK\_ACCT\_ST=3,4 324 44 0 (0.8641975 0.1358025) \*  
## 3) CHK\_ACCT\_ST=1,2 376 153 0 (0.5930851 0.4069149)   
## 6) DUR< 22.5 217 66 0 (0.6958525 0.3041475)   
## 12) CRED\_HIST=3,4,5 197 53 0 (0.7309645 0.2690355)   
## 24) DUR< 11.5 60 7 0 (0.8833333 0.1166667) \*  
## 25) DUR>=11.5 137 46 0 (0.6642336 0.3357664)   
## 50) CRED\_AMT>=1387.5 88 21 0 (0.7613636 0.2386364) \*  
## 51) CRED\_AMT< 1387.5 49 24 1 (0.4897959 0.5102041)   
## 102) PROPERTY=1,2,4 32 10 0 (0.6875000 0.3125000)   
## 204) PURPOSE=3,4,7,8,9 23 4 0 (0.8260870 0.1739130) \*  
## 205) PURPOSE=1,5 9 3 1 (0.3333333 0.6666667) \*  
## 103) PROPERTY=3 17 2 1 (0.1176471 0.8823529) \*  
## 13) CRED\_HIST=1,2 20 7 1 (0.3500000 0.6500000)   
## 26) PURPOSE=2,3,4,8,9 12 5 0 (0.5833333 0.4166667) \*  
## 27) PURPOSE=1 8 0 1 (0.0000000 1.0000000) \*  
## 7) DUR>=22.5 159 72 1 (0.4528302 0.5471698)   
## 14) SAV\_ACCT\_BOND=4,5 24 5 0 (0.7916667 0.2083333) \*  
## 15) SAV\_ACCT\_BOND=1,2,3 135 53 1 (0.3925926 0.6074074)   
## 30) INST\_RT\_PER\_DISP\_INCM< 2.5 52 23 0 (0.5576923 0.4423077)   
## 60) PURPOSE=1,2,4,7,10 30 7 0 (0.7666667 0.2333333) \*  
## 61) PURPOSE=3,9 22 6 1 (0.2727273 0.7272727) \*  
## 31) INST\_RT\_PER\_DISP\_INCM>=2.5 83 24 1 (0.2891566 0.7108434)   
## 62) CRED\_AMT< 3962 50 20 1 (0.4000000 0.6000000)   
## 124) CRED\_AMT>=2313 31 14 0 (0.5483871 0.4516129)   
## 248) EMPLYMT\_ST=1,5 10 2 0 (0.8000000 0.2000000) \*  
## 249) EMPLYMT\_ST=2,3,4 21 9 1 (0.4285714 0.5714286)   
## 498) HOUS\_ST=2 12 4 0 (0.6666667 0.3333333) \*  
## 499) HOUS\_ST=1,3 9 1 1 (0.1111111 0.8888889) \*  
## 125) CRED\_AMT< 2313 19 3 1 (0.1578947 0.8421053) \*  
## 63) CRED\_AMT>=3962 33 4 1 (0.1212121 0.8787879) \*

prp(ct1, type = 1, extra = 1, under = TRUE, split.font = 2, varlen = -10)



pred.val = predict(ct1, data5.val, type = 'class')  
conf(data5.val$Y, pred.val)

## $conf.mat  
## actual  
## pred 0 1  
## 0 176 69  
## 1 21 34  
##   
## $err  
## [1] 0.3  
##   
## $accuracy  
## [1] 0.7  
##   
## $sensitivity  
## [1] 0.3300971  
##   
## $specificity  
## [1] 0.893401

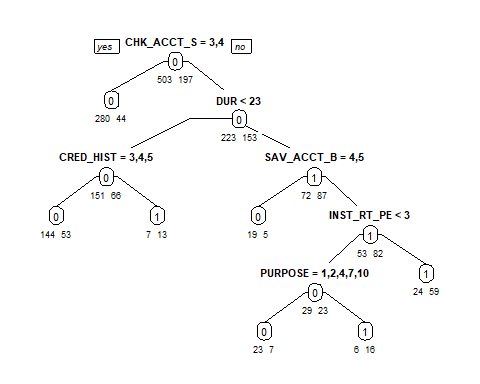
cv.ct <- rpart(Y ~ . , data = data5.train, method = "class", cp = 0.00001, minsplit = 5, xval = 5)  
printcp(cv.ct)

##   
## Classification tree:  
## rpart(formula = Y ~ ., data = data5.train, method = "class",   
## cp = 1e-05, minsplit = 5, xval = 5)  
##   
## Variables actually used in tree construction:  
## [1] AGE CHK\_ACCT\_ST COAPP\_GURNTR   
## [4] CRED\_AMT CRED\_HIST DUR   
## [7] DUR\_RES EMPLYMT\_ST HOUS\_ST   
## [10] INST\_RT\_PER\_DISP\_INCM JOB NUM\_CRED   
## [13] NUM\_PEOP\_LIABL PROPERTY PURPOSE   
## [16] SAV\_ACCT\_BOND   
##   
## Root node error: 197/700 = 0.28143  
##   
## n= 700   
##   
## CP nsplit rel error xerror xstd  
## 1 0.0380711 0 1.00000 1.00000 0.060395  
## 2 0.0304569 5 0.77157 0.91878 0.058804  
## 3 0.0219966 6 0.74112 0.88325 0.058043  
## 4 0.0152284 9 0.67513 0.97462 0.059919  
## 5 0.0126904 10 0.65990 0.98985 0.060207  
## 6 0.0114213 12 0.63452 0.97970 0.060016  
## 7 0.0101523 16 0.58883 0.97970 0.060016  
## 8 0.0076142 28 0.46701 1.01523 0.060672  
## 9 0.0067682 37 0.39086 0.98985 0.060207  
## 10 0.0050761 40 0.37056 1.03046 0.060941  
## 11 0.0033841 60 0.26904 1.05076 0.061291  
## 12 0.0025381 63 0.25888 1.06091 0.061461  
## 13 0.0020305 67 0.24873 1.06091 0.061461  
## 14 0.0016920 74 0.23350 1.05584 0.061376  
## 15 0.0000100 77 0.22843 1.05584 0.061376

prune.ct <- prune(cv.ct, cp = cv.ct$cptable[which.min(cv.ct$cptable[,"xerror"]), "CP"])  
print(prune.ct)

## n= 700   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 700 197 0 (0.7185714 0.2814286)   
## 2) CHK\_ACCT\_ST=3,4 324 44 0 (0.8641975 0.1358025) \*  
## 3) CHK\_ACCT\_ST=1,2 376 153 0 (0.5930851 0.4069149)   
## 6) DUR< 22.5 217 66 0 (0.6958525 0.3041475)   
## 12) CRED\_HIST=3,4,5 197 53 0 (0.7309645 0.2690355) \*  
## 13) CRED\_HIST=1,2 20 7 1 (0.3500000 0.6500000) \*  
## 7) DUR>=22.5 159 72 1 (0.4528302 0.5471698)   
## 14) SAV\_ACCT\_BOND=4,5 24 5 0 (0.7916667 0.2083333) \*  
## 15) SAV\_ACCT\_BOND=1,2,3 135 53 1 (0.3925926 0.6074074)   
## 30) INST\_RT\_PER\_DISP\_INCM< 2.5 52 23 0 (0.5576923 0.4423077)   
## 60) PURPOSE=1,2,4,7,10 30 7 0 (0.7666667 0.2333333) \*  
## 61) PURPOSE=3,9 22 6 1 (0.2727273 0.7272727) \*  
## 31) INST\_RT\_PER\_DISP\_INCM>=2.5 83 24 1 (0.2891566 0.7108434) \*

prp(prune.ct, type = 1, extra = 1, under = TRUE, split.font = 2, varlen = -10)



pred.val = predict(prune.ct, data5.val, type = 'class')  
conf(data5.val$Y, pred.val)

## $conf.mat  
## actual  
## pred 0 1  
## 0 178 66  
## 1 19 37  
##   
## $err  
## [1] 0.2833333  
##   
## $accuracy  
## [1] 0.7166667  
##   
## $sensitivity  
## [1] 0.3592233  
##   
## $specificity  
## [1] 0.9035533

dt = predict(prune.ct,data5.val,type = "prob",probability =TRUE)  
auc(ifelse(data5.val$Y == "0",1, 0), dt[,"1"])

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

## Area under the curve: 0.7364