Data Mining (IDS 472)
HOMEWORK 3

Oleh Antonyuk
Ravi Patel
Warren Svoboda

Problem 1.

A.)

```
x <- salary_class

#converts into factors.
x$INCOME = factor(s$INCOME)
x$EMPLOYER = factor(s$EMPLOYER)
x$DEGREE = factor(s$DEGREE)
x$MSTATUS = factor(s$MSTATUS)
x$JOBTYPE = factor(s$JOBTYPE)
x$SEX = factor(s$SEX)
x$COUNTRY = factor(s$COUNTRY)

describe(s)
dim(s)

str(s)

#cleaning missing values
zgain = (s$`C-GAIN` - mean(s$`C-GAIN`))/sd(x$`C-GAIN`)
sum(abs(zgain) > 1)
sum(abs(zgain) < -1)

x$`C-GAIN` = ifelse(abs(zgain) > 1, NA, x$`C-GAIN`)
x$`C-GAIN` = ifelse(abs(zgain) < -1, NA, x$`C-GAIN`)

zloss = (x$`C-LOSS` - mean(x$`C-LOSS`))/sd(x$`C-LOSS`)
sum(abs(zloss) > 1)
sum(abs(zloss) < -1)

x$`C-LOSS` = ifelse(abs(zloss) > 1, NA, x$`C-LOSS`)
x$`C-LOSS` = ifelse(abs(zloss) < -1, NA, x$`C-LOSS`)

#deletes all the NA values from CGAIN and CLOSS
na.omit(x$`C-GAIN`)
na.omit(x$`C-LOSS`)
```

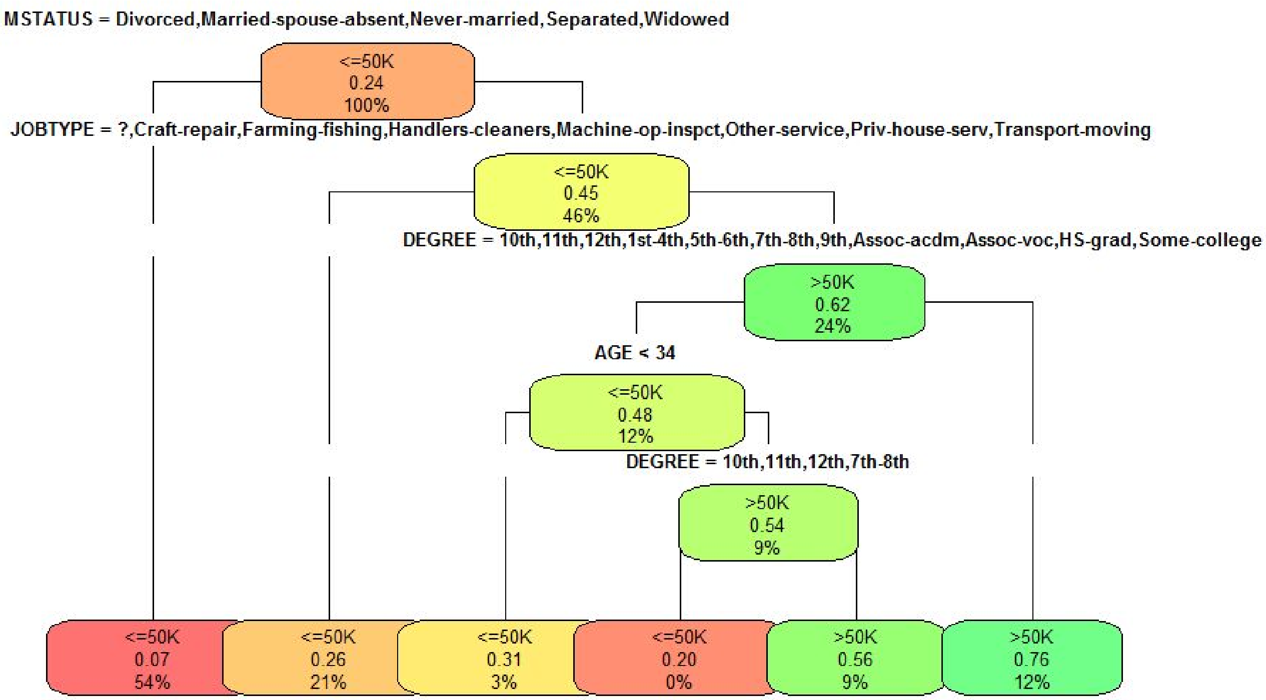
B)

```
set.seed(1234)
index = sample(2, nrow(s), replace = T, prob = c(.6, .4))
trainData = x[index == 1,]
testData = x[index == 2,]
summary(x)
```

```
myFormula = INCOME ~ AGE + DEGREE + EMPLOYER + MSTATUS + JOBTYPE + SEX +
                `C-LOSS` + `C-GAIN` + COUNTRY + INCOME
PARSHIMANrtree = rpart(myFormula, data=trainData)
rpart.plot(PARSHIMANrtree,type = 1,
        box.palette = "auto")
prp(PARSHIMANrtree, faclen = 0, cex = .6, extra = 1)
summary(PARSHIMANrtree)


asRules(PARSHIMANrtree)

printcp(PARSHIMANrtree)
```



6 leaves

C) With many input variables, the most effective variable in predicting income are marital status (MSTATUS) and job type (JOBTYPE). This comes from the "summary(PARSHIMANrtree)" function.

D)

```
install.packages("rattle")
library(rattle)
```

```
asRules(PARSHIMANrtree)
Rule number: 15 [INCOME=>50K cover=2331 (12%) prob=0.76]
  MSTATUS=Married-AF-spouse,Married-civ-spouse
  JOBTYPE=Adm-clerical,Armed-Forces,Exec-managerial,Prof-specialty,Protective-serv,Sales,Tech-support
  DEGREE=Bachelors,Doctorate,Masters,Prof-school

 Rule number: 59 [INCOME=>50K cover=1752 (9%) prob=0.56]
  MSTATUS=Married-AF-spouse,Married-civ-spouse
  JOBTYPE=Adm-clerical,Armed-Forces,Exec-managerial,Prof-specialty,Protective-serv,Sales,Tech-support
  DEGREE=10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Some-college
  AGE>=33.5
  DEGREE=1st-4th,5th-6th,9th,Assoc-acdm,Assoc-voc,HS-grad,Some-college


 Rule number: 6 [INCOME=<=50K cover=4151 (21%) prob=0.26]
  MSTATUS=Married-AF-spouse,Married-civ-spouse

JOBTYPE=?,Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Transp
ort-moving

 Rule number: 2 [INCOME=<=50K cover=10553 (54%) prob=0.07]
  MSTATUS=Divorced,Married-spouse-absent,Never-married,Separated,Widowed
```


```
Rule number: 6 [INCOME=<=50K cover=4151 (21%) prob=0.26]
  MSTATUS=Married-AF-spouse,Married-civ-spouse

JOBTYPE=?,Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Transp
ort-moving
```


If you want to earn the >50K income category, married, spouses at these jobs = armed forces, or some executive/upper mgmt position with age

Characteristic of those with Income <=50K

- Not in a standard (typical) marriage
- Service/physical labor intensive jobs (Crafting, Agriculture)
- Less than 34 years of age

Characteristic of those with Income >50K

- Married
- Often in an executive/upper management position
- Greater than 33.5 years of age

E)

```
t = predict(PARSHIMANrtree ,type= "class", newdata = testData)
table(t, testData$INCOME)
summary(PARSHIMANrtree)

PARSHIMANrtree= rpart(myFormula, data = trainData, parms = list(split= "gini"))
rpart.plot(PARSHIMANrtree,type = 1,box.palette = "auto")
```

P0ARSHIMANrtrees= rpart(myFormula, data = trainData, parms = list(split= "gini"), cp = 0.0005,
control = rpart.control(minsplit = 500, minbucket = 100))

rpart.plot(PARSHIMANrtrees,type = 1,
box.palette = "auto")
plot(PARSHIMANrtrees)
text(PARSHIMANrtrees, use.n = T, xpd = T)
PARSHIMANrtrees$cptable

In general, the data has higher accuracy after pruning. With the un-pruned data yielding greater accuracy on the training data, and the test data with higher accuracy on the pruned data.

Problem 2.

2a)

Code for the Setup and Train Data CP = 0.005

```
BankLoan$checking_balance=factor(BankLoan$checking_balance)
BankLoan$credit_history=factor(BankLoan$credit_history)
BankLoan$purpose=factor(BankLoan$purpose)
BankLoan$savings_balance=factor(BankLoan$savings_balance)
BankLoan$employment_length=factor(BankLoan$employment_length)
BankLoan$other_debtors=factor(BankLoan$other_debtors)
BankLoan$personal_status=factor(BankLoan$personal_status)
BankLoan$property=factor(BankLoan$property)
BankLoan$installment_plan=factor(BankLoan$installment_plan)
BankLoan$housing=factor(BankLoan$housing)
BankLoan$telephone=factor(BankLoan$telephone)
BankLoan$foreign_worker=factor(BankLoan$foreign_worker)
BankLoan$job=factor(BankLoan$job)
BankLoan$default=factor(BankLoan$default)
BankLoan$installment_rate=factor(BankLoan$installment_rate)
BankLoan$residence_history=factor(BankLoan$residence_history)
BankLoan$age=factor(BankLoan$age)
BankLoan$existing_credits=factor(BankLoan$existing_credits)
BankLoan$dependents=factor(BankLoan$dependents)
set.seed(1234)
```

```
> index = sample(2, nrow(BankLoan), replace = T, prob = c(0.7,0.3))
> TrainData = BankLoan[index == 1, ]
> TestData = BankLoan[index == 2,]
> library(rpart)
> bl_rpart = rpart(default~., data = TrainData, method = "class", control = rpart.control(cp = 0.005), parms
 = list(split = "information"))
> prediction.matrix<-table(predict(bl_rpart, newdata = TrainData, type = "class"), TrainData$default)
> prediction.matrix

      1   2
  1 459  56
  2  29 157
> accuracy<-sum(diag(prediction.matrix))/sum(prediction.matrix)
> accuracy
[1] 0.8787447
```

Train Data has higher accuracy

Test Data Numbers CP = 0.005

```
> prediction.matrix<-table(predict(bl_rpart, newdata = TestData, type = "class"), TestData$default)
> prediction.matrix

      1   2
  1 167  55
  2  45  32
> accuracy<-sum(diag(prediction.matrix))/sum(prediction.matrix)
> accuracy
[1] 0.6655518
> 
```

```
> library(rpart)
> bl_rpart
n= 701

node), split, n, loss, yval, (yprob)
      * denotes terminal node

    1) root 701 213 1 (0.69614836 0.30385164)
      2) checking_balance=unknown 276  31 1 (0.88768116 0.11231884)
        4) age=20,34,40,41,42,45,46,48,49,50,51,52,54,56,57,59,60,61,62,63,64,65,74 80   0 1
(1.00000000 0.00000000) *
        5) age=19,21,22,23,24,25,26,27,28,29,30,31,32,33,35,36,37,38,39,43,44,47,53,55,68 196
31 1 (0.84183673 0.15816327)
         10) installment_plan=none,stores 171  18 1 (0.89473684 0.10526316)
           20) age=24,27,29,33,35,38,39,44,53 60   0 1 (1.00000000 0.00000000) *
           21) age=19,21,22,23,25,26,28,30,31,32,36,37,43,47,55,68 111  18 1 (0.83783784
0.16216216)
             42) purpose=car (used),domestic appliances,radio/tv,retraining 51   2 1 (0.96078431
0.03921569) *
             43) purpose=business,car (new),education,furniture,repairs 60  16 1 (0.73333333
0.26666667)
               86) age=21,22,23,25,26,30,31,32,36,37,43 52   9 1 (0.82692308 0.17307692) *
               87) age=19,28,47,55,68 8   1 2 (0.12500000 0.87500000) *
         11) installment_plan=bank 25  12 2 (0.48000000 0.52000000)
           22) purpose=furniture,radio/tv 12   3 1 (0.75000000 0.25000000) *
           23) purpose=business,car (new),car (used) 13   3 2 (0.23076923 0.76923077) *
      3) checking_balance=< 0 DM,> 200 DM,1 - 200 DM 425 182 1 (0.57176471 0.42823529)
        6) age=19,20,27,30,32,35,36,37,38,40,41,44,45,48,49,51,54,63,64,66,67,75 179  48 1
(0.73184358 0.26815642)
         12) credit_history=critical,delayed 64   8 1 (0.87500000 0.12500000) *
         13) credit_history=fully repaid,fully repaid this bank,repaid 115  40 1 (0.65217391
0.34782609)
           26) property=building society savings,other,real estate 93  26 1 (0.72043011
0.27956989)
             52) age=19,38,40,48,49,64,66,67,75 19   0 1 (1.00000000 0.00000000) *
             53) age=20,27,30,32,35,36,37,41,44,45,51,54 74  26 1 (0.64864865 0.35135135)
              106) months_loan_duration< 8 8   0 1 (1.00000000 0.00000000) *
              107) months_loan_duration>=8 66  26 1 (0.60606061 0.39393939)
                214) savings_balance=> 1000 DM,501 - 1000 DM 7   0 1 (1.00000000 0.00000000) *
                215) savings_balance=< 100 DM,101 - 500 DM,unknown 59  26 1 (0.55932203
0.44067797)
                  430) amount>=3261 22   5 1 (0.77272727 0.22727273) *
```

```
            431) amount< 3261 37  16 2 (0.43243243 0.56756757)
              862) purpose=business,car (used),domestic appliances,radio/tv,retraining 15   4 1
(0.73333333 0.26666667) *
              863) purpose=car (new),education,furniture,repairs 22   5 2 (0.22727273
0.77272727) *
         27) property=unknown/none 22   8 2 (0.36363636 0.63636364)
           54) age=30,36,40,51,63 11   3 1 (0.72727273 0.27272727) *
           55) age=27,32,35,38,44,48 11   0 2 (0.00000000 1.00000000) *
      7)
age=21,22,23,24,25,26,28,29,31,33,34,39,42,43,46,47,50,52,53,55,57,58,59,60,61,65,68,74
246 112 2 (0.45528455 0.54471545)
       14) months_loan_duration< 31.5 191  89 1 (0.53403141 0.46596859)
         28) purpose=business,car (used),education,furniture,radio/tv 130  50 1 (0.61538462
0.38461538)
           56) age=21,22,23,24,25,26,28,29,31,33,34,42,43,47,50,55,57,59,60,61 122  43 1
(0.64754098 0.35245902)
            112) age=21,28,47,50,59 17   2 1 (0.88235294 0.11764706) *
            113) age=22,23,24,25,26,29,31,33,34,42,43,55,57,60,61 105  41 1 (0.60952381
0.39047619)
              226) purpose=car (used) 7   0 1 (1.00000000 0.00000000) *
              227) purpose=business,education,furniture,radio/tv 98  41 1 (0.58163265
0.41836735)
                454) savings_balance=> 1000 DM,501 - 1000 DM,unknown 24   5 1 (0.79166667
0.20833333) *
                455) savings_balance=< 100 DM,101 - 500 DM 74  36 1 (0.51351351 0.48648649)
                  910) credit_history=critical,delayed,repaid 61  25 1 (0.59016393 0.40983607)
                   1820) other_debtors=guarantor 8   0 1 (1.00000000 0.00000000) *
                   1821) other_debtors=co-applicant,none 53  25 1 (0.52830189 0.47169811)
                     3642) months_loan_duration< 16.5 25   7 1 (0.72000000 0.28000000) *
                     3643) months_loan_duration>=16.5 28  10 2 (0.35714286 0.64285714)
                       7286) age=22,23,24,26,33,34,43,55 20  10 1 (0.50000000 0.50000000)
                        14572) personal_status=divorced male,single male 8   2 1 (0.75000000
0.25000000) *
                        14573) personal_status=female,married male 12   4 2 (0.33333333
0.66666667) *
                       7287) age=25,29,31,42,61 8   0 2 (0.00000000 1.00000000) *
                  911) credit_history=fully repaid,fully repaid this bank 13   2 2 (0.15384615
0.84615385) *
           57) age=39,52,53,65,74 8   1 2 (0.12500000 0.87500000) *
         29) purpose=car (new),domestic appliances,others,repairs,retraining 61  22 2
(0.36065574 0.63934426)
           58) age=22,23,24,25,26,28,29,31,33,39,42,47,58,65 49  22 2 (0.44897959 0.55102041)
            116) age=26,31,39,42,58,65 12   3 1 (0.75000000 0.25000000) *
```

117) age=22,23,24,25,28,29,33,47 37 13 2 (0.35135135 0.64864865)
234) amount>=1387 20 9 1 (0.55000000 0.45000000)
468) existing_credits=1 13 3 1 (0.76923077 0.23076923) *
469) existing_credits=2,3 7 1 2 (0.14285714 0.85714286) *
235) amount< 1387 17 2 2 (0.11764706 0.88235294) *
59) age=21,34,43,46,53,55,60,61,68 12 0 2 (0.00000000 1.00000000) *
15) months_loan_duration>=31.5 55 10 2 (0.18181818 0.81818182) *

2b) checking_balance is the biggest determinant because it is the first split in the decision tree.

Party tree variation

Model formula:
default ~ checking_balance + months_loan_duration + credit_history +
purpose + amount + savings_balance + employment_length +
installment_rate + personal_status + other_debtors + residence_history +
property + age + installment_plan + housing + existing_credits +
dependents + telephone + foreign_worker + job

Fitted party:
[1] root
| [2] checking_balance in unknown
| | [3] age in 20, 34, 40, 41, 42, 45, 46, 48, 49, 50, 51, 52, 54, 56, 57, 59, 60, 61, 62, 63, 64, 65, 74: 1 (n = 80, err = 0.0%)
| | [4] age in 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 43, 44, 47, 53, 55, 68
| | | [5] installment_plan in none, stores
| | | | [6] age in 24, 27, 29, 33, 35, 38, 39, 44, 53: 1 (n = 60, err = 0.0%)
| | | | [7] age in 19, 21, 22, 23, 25, 26, 28, 30, 31, 32, 36, 37, 43, 47, 55, 68
| | | | | [8] purpose in car (used), domestic appliances, radio/tv, retraining: 1 (n = 51, err = 3.9%)
| | | | | [9] purpose in business, car (new), education, furniture, repairs
| | | | | | [10] age in 21, 22, 23, 25, 26, 30, 31, 32, 36, 37, 43: 1 (n = 52, err = 17.3%)
| | | | | | [11] age in 19, 28, 47, 55, 68: 2 (n = 8, err = 12.5%)
| | | [12] installment_plan in bank
| | | | [13] purpose in furniture, radio/tv: 1 (n = 12, err = 25.0%)
| | | | [14] purpose in business, car (new), car (used): 2 (n = 13, err = 23.1%)
| [15] checking_balance < 0 DM, > 200 DM, 1 - 200 DM
| | [16] age in 19, 20, 27, 30, 32, 35, 36, 37, 38, 40, 41, 44, 45, 48, 49, 51, 54, 63, 64, 66, 67, 75
| | | [17] credit_history in critical, delayed: 1 (n = 64, err = 12.5%)
| | | [18] credit_history in fully repaid, fully repaid this bank, repaid
| | | | [19] property in building society savings, other, real estate
| | | | | [20] age in 19, 38, 40, 48, 49, 64, 66, 67, 75: 1 (n = 19, err = 0.0%)
| | | | | [21] age in 20, 27, 30, 32, 35, 36, 37, 41, 44, 45, 51, 54
| | | | | | [22] months_loan_duration < 8: 1 (n = 8, err = 0.0%)

| | | | | | [23] months_loan_duration >= 8
| | | | | | | [24] savings_balance > 1000 DM, 501 - 1000 DM: 1 (n = 7, err = 0.0%)
| | | | | | | [25] savings_balance < 100 DM, 101 - 500 DM, unknown
| | | | | | | | [26] amount >= 3261: 1 (n = 22, err = 22.7%)
| | | | | | | | [27] amount < 3261
| | | | | | | | | [28] purpose in business, car (used), domestic appliances, radio/tv, retraining: 1 (n = 15, err = 26.7%)
| | | | | | | | | [29] purpose in car (new), education, furniture, repairs: 2 (n = 22, err = 22.7%)
| | | | [30] property in unknown/none
| | | | | [31] age in 30, 36, 40, 51, 63: 1 (n = 11, err = 27.3%)
| | | | | [32] age in 27, 32, 35, 38, 44, 48: 2 (n = 11, err = 0.0%)
| | [33] age in 21, 22, 23, 24, 25, 26, 28, 29, 31, 33, 34, 39, 42, 43, 46, 47, 50, 52, 53, 55, 57, 58, 59, 60, 61, 65, 68, 74
| | | [34] months_loan_duration < 31.5
| | | | [35] purpose in business, car (used), education, furniture, radio/tv
| | | | | [36] age in 21, 22, 23, 24, 25, 26, 28, 29, 31, 33, 34, 42, 43, 47, 50, 55, 57, 59, 60, 61
| | | | | | [37] age in 21, 28, 47, 50, 59: 1 (n = 17, err = 11.8%)
| | | | | | [38] age in 22, 23, 24, 25, 26, 29, 31, 33, 34, 42, 43, 55, 57, 60, 61
| | | | | | | [39] purpose in car (used): 1 (n = 7, err = 0.0%)
| | | | | | | [40] purpose in business, education, furniture, radio/tv
| | | | | | | | [41] savings_balance > 1000 DM, 501 - 1000 DM, unknown: 1 (n = 24, err = 20.8%)
| | | | | | | | [42] savings_balance < 100 DM, 101 - 500 DM
| | | | | | | | | [43] credit_history in critical, delayed, repaid
| | | | | | | | | | [44] other_debtors in guarantor: 1 (n = 8, err = 0.0%)
| | | | | | | | | | [45] other_debtors in co-applicant, none
| | | | | | | | | | | [46] months_loan_duration < 16.5: 1 (n = 25, err = 28.0%)
| | | | | | | | | | | [47] months_loan_duration >= 16.5
| | | | | | | | | | | | [48] age in 22, 23, 24, 26, 33, 34, 43, 55
| | | | | | | | | | | | | [49] personal_status in divorced male, single male: 1 (n = 8, err = 25.0%)
| | | | | | | | | | | | | [50] personal_status in female, married male: 2 (n = 12, err = 33.3%)
| | | | | | | | | | | | [51] age in 25, 29, 31, 42, 61: 2 (n = 8, err = 0.0%)
| | | | | | | | | [52] credit_history in fully repaid, fully repaid this bank: 2 (n = 13, err = 15.4%)
| | | | | [53] age in 39, 52, 53, 65, 74: 2 (n = 8, err = 12.5%)
| | | | [54] purpose in car (new), domestic appliances, others, repairs, retraining
| | | | | [55] age in 22, 23, 24, 25, 26, 28, 29, 31, 33, 39, 42, 47, 58, 65
| | | | | | [56] age in 26, 31, 39, 42, 58, 65: 1 (n = 12, err = 25.0%)
| | | | | | [57] age in 22, 23, 24, 25, 28, 29, 33, 47

| | | | | | | [58] amount >= 1387
| | | | | | | | [59] existing_credits in 1: 1 (n = 13, err = 23.1%)
| | | | | | | | [60] existing_credits in 2, 3: 2 (n = 7, err = 14.3%)
| | | | | | | [61] amount < 1387: 2 (n = 17, err = 11.8%)
| | | | | [62] age in 21, 34, 43, 46, 53, 55, 60, 61, 68: 2 (n = 12, err = 0.0%)
| | | [63] months_loan_duration >= 31.5: 2 (n = 55, err = 18.2%)

Number of inner nodes: 31
Number of terminal nodes: 32

```
> predictdata<-data.frame(checking_balance=c("1 - 200 DM"), credit_history=c("delayed"), purpose=c("furnitur
e"), amount=c(1913), savings_balance=c("< 100 DM"), employment_length=c("1 - 4 yrs"), age=c("48"),  housing=
c("rent"), existing_credits=c("2"), telephone=c("yes"), foreign_worker=c("yes"),  job=c("skilled employee"),
 dependents = c("1"), residence_history = c("4"), installment_rate = c("4"), personal_status = c("single mal
e"), installment_plan = c("none"), other_debtors = c("none"), months_loan_duration = c(20.9), property = c("
other"))
>
> predict(bl_rpart,predictdata)
      1     2
1 0.875 0.125
> 
```

2c) P(default) is 0.875 in the scenario where 1=no default and 2=defaulted
> as.party(bl_rpart)

Model formula:
default ~ checking_balance + months_loan_duration + credit_history +
purpose + amount + savings_balance + employment_length +
installment_rate + personal_status + other_debtors + residence_history +
property + age + installment_plan + housing + existing_credits +
dependents + telephone + foreign_worker + job

Fitted party:
[1] root
| [2] checking_balance in unknown
| | [3] age in 20, 34, 40, 41, 42, 45, 46, 48, 49, 50, 51, 52, 54, 56, 57, 59, 60, 61, 62, 63, 64, 65, 74: 1 (n = 80, err = 0.0%)
| | [4] age in 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 43, 44, 47, 53, 55, 68
| | | [5] installment_plan in none, stores
| | | | [6] age in 24, 27, 29, 33, 35, 38, 39, 44, 53: 1 (n = 60, err = 0.0%)
| | | | [7] age in 19, 21, 22, 23, 25, 26, 28, 30, 31, 32, 36, 37, 43, 47, 55, 68
| | | | | [8] purpose in car (used), domestic appliances, radio/tv, retraining: 1 (n = 51, err = 3.9%)
| | | | | [9] purpose in business, car (new), education, furniture, repairs
| | | | | | [10] age in 21, 22, 23, 25, 26, 30, 31, 32, 36, 37, 43: 1 (n = 52, err = 17.3%)
| | | | | | [11] age in 19, 28, 47, 55, 68: 2 (n = 8, err = 12.5%)
| | | [12] installment_plan in bank
| | | | [13] purpose in furniture, radio/tv: 1 (n = 12, err = 25.0%)

| | | | [14] purpose in business, car (new), car (used): 2 (n = 13, err = 23.1%)
| [15] checking_balance < 0 DM, > 200 DM, 1 - 200 DM
| | [16] age in 19, 20, 27, 30, 32, 35, 36, 37, 38, 40, 41, 44, 45, 48, 49, 51, 54, 63, 64, 66, 67, 75
| | | [17] credit_history in critical, delayed: 1 (n = 64, err = 12.5%)
| | | [18] credit_history in fully repaid, fully repaid this bank, repaid
| | | | [19] property in building society savings, other, real estate
| | | | | [20] age in 19, 38, 40, 48, 49, 64, 66, 67, 75: 1 (n = 19, err = 0.0%)
| | | | | [21] age in 20, 27, 30, 32, 35, 36, 37, 41, 44, 45, 51, 54
| | | | | | [22] months_loan_duration < 8: 1 (n = 8, err = 0.0%)
| | | | | | [23] months_loan_duration >= 8
| | | | | | | [24] savings_balance > 1000 DM, 501 - 1000 DM: 1 (n = 7, err = 0.0%)
| | | | | | | [25] savings_balance < 100 DM, 101 - 500 DM, unknown
| | | | | | | | [26] amount >= 3261: 1 (n = 22, err = 22.7%)
| | | | | | | | [27] amount < 3261
| | | | | | | | | [28] purpose in business, car (used), domestic appliances, radio/tv, retraining: 1 (n = 15, err = 26.7%)
| | | | | | | | | [29] purpose in car (new), education, furniture, repairs: 2 (n = 22, err = 22.7%)
| | | | [30] property in unknown/none
| | | | | [31] age in 30, 36, 40, 51, 63: 1 (n = 11, err = 27.3%)
| | | | | [32] age in 27, 32, 35, 38, 44, 48: 2 (n = 11, err = 0.0%)
| | [33] age in 21, 22, 23, 24, 25, 26, 28, 29, 31, 33, 34, 39, 42, 43, 46, 47, 50, 52, 53, 55, 57, 58, 59, 60, 61, 65, 68, 74
| | | [34] months_loan_duration < 31.5
| | | | [35] purpose in business, car (used), education, furniture, radio/tv
| | | | | [36] age in 21, 22, 23, 24, 25, 26, 28, 29, 31, 33, 34, 42, 43, 47, 50, 55, 57, 59, 60, 61
| | | | | | [37] age in 21, 28, 47, 50, 59: 1 (n = 17, err = 11.8%)
| | | | | | [38] age in 22, 23, 24, 25, 26, 29, 31, 33, 34, 42, 43, 55, 57, 60, 61
| | | | | | | [39] purpose in car (used): 1 (n = 7, err = 0.0%)
| | | | | | | [40] purpose in business, education, furniture, radio/tv
| | | | | | | | [41] savings_balance > 1000 DM, 501 - 1000 DM, unknown: 1 (n = 24, err = 20.8%)
| | | | | | | | [42] savings_balance < 100 DM, 101 - 500 DM
| | | | | | | | | [43] credit_history in critical, delayed, repaid
| | | | | | | | | | [44] other_debtors in guarantor: 1 (n = 8, err = 0.0%)
| | | | | | | | | | [45] other_debtors in co-applicant, none
| | | | | | | | | | | [46] months_loan_duration < 16.5: 1 (n = 25, err = 28.0%)
| | | | | | | | | | | [47] months_loan_duration >= 16.5
| | | | | | | | | | | | [48] age in 22, 23, 24, 26, 33, 34, 43, 55
| | | | | | | | | | | | | [49] personal_status in divorced male, single male: 1 (n = 8, err = 25.0%)

| | | | | | | | | | | | | [50] personal_status in female, married male: 2 (n = 12, err = 33.3%)
| | | | | | | | | | | | [51] age in 25, 29, 31, 42, 61: 2 (n = 8, err = 0.0%)
| | | | | | | | | [52] credit_history in fully repaid, fully repaid this bank: 2 (n = 13, err = 15.4%)
| | | | | [53] age in 39, 52, 53, 65, 74: 2 (n = 8, err = 12.5%)
| | | | [54] purpose in car (new), domestic appliances, others, repairs, retraining
| | | | | [55] age in 22, 23, 24, 25, 26, 28, 29, 31, 33, 39, 42, 47, 58, 65
| | | | | | [56] age in 26, 31, 39, 42, 58, 65: 1 (n = 12, err = 25.0%)
| | | | | | [57] age in 22, 23, 24, 25, 28, 29, 33, 47
| | | | | | | [58] amount >= 1387
| | | | | | | | [59] existing_credits in 1: 1 (n = 13, err = 23.1%)
| | | | | | | | [60] existing_credits in 2, 3: 2 (n = 7, err = 14.3%)
| | | | | | | [61] amount < 1387: 2 (n = 17, err = 11.8%)
| | | | | [62] age in 21, 34, 43, 46, 53, 55, 60, 61, 68: 2 (n = 12, err = 0.0%)
| | | [63] months_loan_duration >= 31.5: 2 (n = 55, err = 18.2%)

Number of inner nodes: 31
Number of terminal nodes: 32

2d) The customer least likely to pay is that with the greatest error. Here we have node 50 with an error of 33.3%. They have personal_status as female and married male. They oftentimes have no current employment and low credit history.

2e) The customer most likely to repay their loan is that with the least error. There are several nodes that have 0% error. This customer has installment_plan in stores or none, a low duration on loan of less than eight months, a savings balance of >1000 DM, borrowing money for a used car, and a guarantor as an other debtor.

2f) Train Data Numbers CP = 0.05

```
> blnouveau_rpart = rpart(default~., data = TrainData, method = "class", control = rpart.control(cp = 0.05), parms = list(split = "information"))
>
> as.party(blnouveau_rpart)

Model formula:
default ~ checking_balance + months_loan_duration + credit_history +
    purpose + amount + savings_balance + employment_length +
    installment_rate + personal_status + other_debtors + residence_history +
    property + age + installment_plan + housing + existing_credits +
    dependents + telephone + foreign_worker + job

Fitted party:
[1] root
|   [2] checking_balance in unknown: 1 (n = 276, err = 11.2%)
|   [3] checking_balance < 0 DM, > 200 DM, 1 - 200 DM
|   |   [4] age in 19, 20, 27, 30, 32, 35, 36, 37, 38, 40, 41, 44, 45, 48, 49, 51, 54, 63, 64, 66, 67, 75: 1 (n = 179, err = 26.8%)
|   |   [5] age in 21, 22, 23, 24, 25, 26, 28, 29, 31, 33, 34, 39, 42, 43, 46, 47, 50, 52, 53, 55, 57, 58, 59, 60, 61, 65, 68, 74
|   |   |   [6] months_loan_duration < 31.5
|   |   |   |   [7] purpose in business, car (used), education, furniture, radio/tv: 1 (n = 130, err = 38.5%)
|   |   |   |   [8] purpose in car (new), domestic appliances, others, repairs, retraining: 2 (n = 61, err = 36.1%)
|   |   |   [9] months_loan_duration >= 31.5: 2 (n = 55, err = 18.2%)

Number of inner nodes:    4
Number of terminal nodes: 5
> prediction.matrix<-table(predict(blnouveau_rpart, newdata = TrainData, type = "class"), TrainData$default)
```

```
> prediction.matrix

      1   2
  1 456 129
  2  32  84
> accuracy<-sum(diag(prediction.matrix))/sum(prediction.matrix)
> accuracy
[1] 0.7703281
```

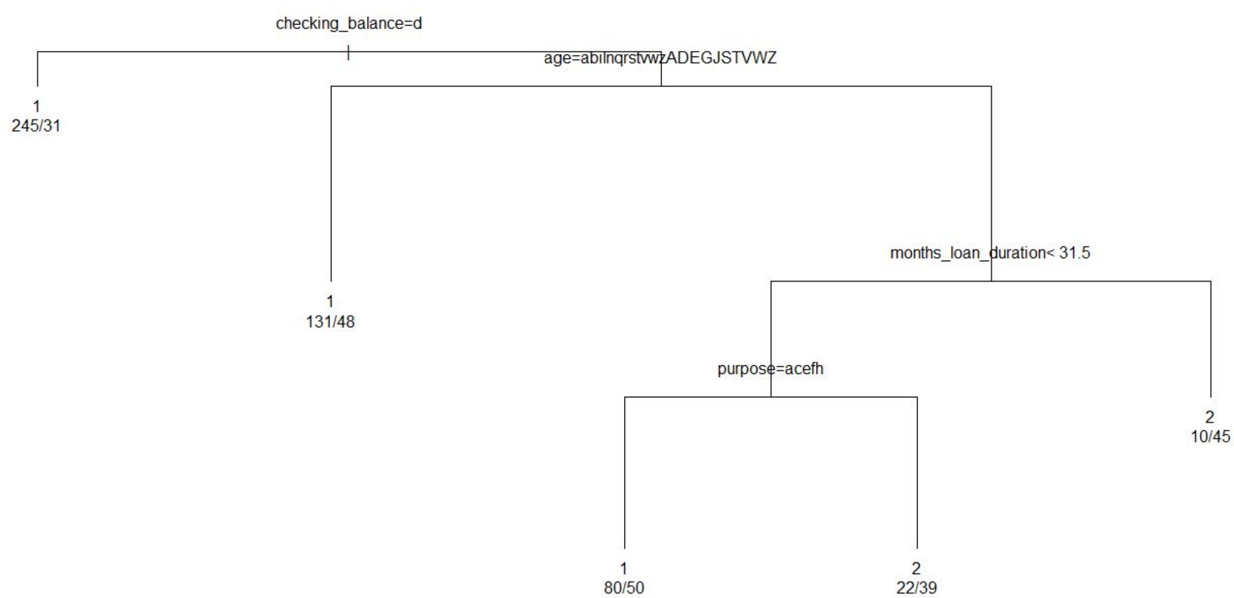
Test Data Numbers CP = 0.05

```
> prediction.matrix<-table(predict(blnouveau_rpart, newdata = TestData, type = "class"), TestData$default)
> prediction.matrix

      1   2
  1 184  61
  2  28  26
> accuracy<-sum(diag(prediction.matrix))/sum(prediction.matrix)
> accuracy
[1] 0.7023411
```

The Train Data has greater accuracy with CP = 0.05
(Train CP = 0.005 Accuracy = .8787) (Test CP = 0.005 Accuracy = .6656)
(Train CP = 0.05 Accuracy = .7703) (Test CP = 0.05 Accuracy = .7023)

2g)

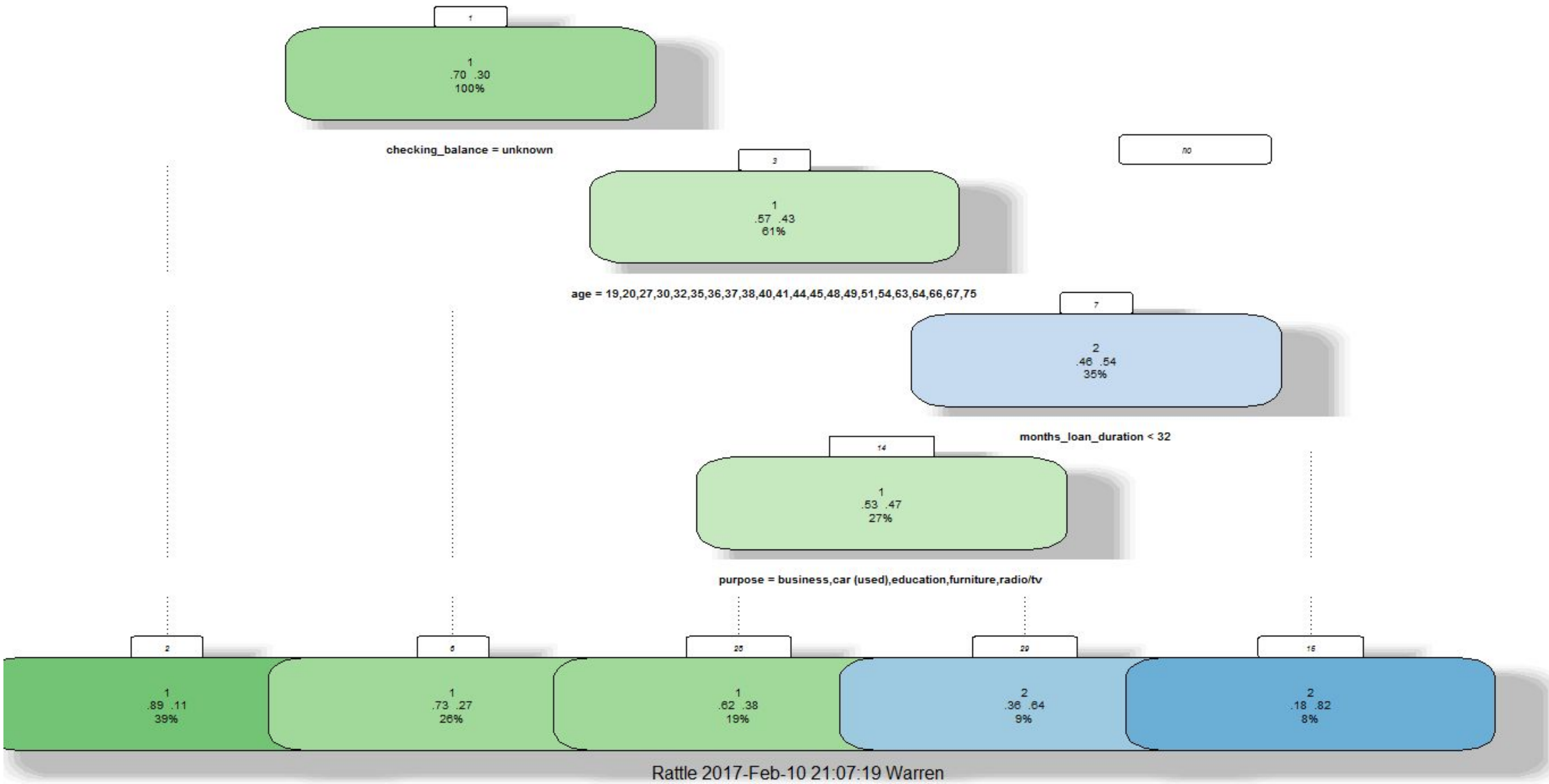


The new decision tree has 5 leaf nodes

2h) The second model does not convey as much information as the first model in terms of predicting ability to repay due to the error at the terminal (leaf) nodes being higher (greater total error = worse prediction power). With a higher CP value, error is likely to increase.

2i) This model predicts that the customer is 73% likely to NOT default on their loan.

```
> predictdatanouveau<-data.frame(checking_balance=c("1 - 200 DM"), credit_history=c("delayed"), purpose=c("furniture"), amount=c(1913), savings_balance=c("< 100 DM"), employment_length=c("1 - 4 yrs"), age=c("48"),  housing=c("rent"), existing_credits=c("2"), telephone=c("yes"), foreign_worker=c("yes"),  job=c("skilled employee"), dependents = c("1"), residence_history = c("4"), installment_rate = c("4"), personal_status = c("single male"), installment_plan = c("none"), other_debtors = c("none"), months_loan_duration = c(20.9), property = c("other"))
>
> predict(blnouveau_rpart,predictdatanouveau)
          1         2
1 0.7318436 0.2681564
```



2j) A higher checking balance will increase the likelihood that a customer will repay their loan.