FIN6120 Credit Scoring Card

Haoyue Heather Tan

3/31/2022

Insert data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- read.csv("credit_data.csv")
#glimpse(df)
#summary(df)
```

```
library(glmmTMB)
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.4.1
## Current Matrix version is 1.3.4
## Please re-install 'TMB' from source using install.packages('TMB', type = 'source') or ask CRAN for a
```

```
## Warning in checkDepPackageVersion(dep_pkg = "TMB"): Package version inconsistency detected.
## glmmTMB was built with TMB version 1.8.0
## Current TMB version is 1.8.1
## Please re-install glmmTMB from source or restore original 'TMB' package (see '?reinstalling' for mor
```

```
y <- as.factor(df$Creditability)
accbl <- as.factor(df$Account.Balance)
pmt <- as.factor(df$Payment.Status.of.Previous.Credit)
value <- as.factor(df$Value.Savings.Stocks)
cEmp <- as.factor(df$Length.of.current.employment)
instpct <- as.factor(df$Instalment.per.cent)
marrital <- as.factor(df$Sex...Marital.Status)
mValue <- as.factor(df$Most.valuable.available.asset)
age <- df$Age..years.
cCre <- as.factor(df$Concurrent.Credits)
foreign <- as.factor(df$Foreign.Worker)
library(dplyr)
duration <- ntile(df$Duration.of.Credit..month, 4)
```

```
#duration <- df$Duration.of.Credit..month.
df <- data.frame(y, accbl,pmt, value,cEmp, instpct, marrital,mValue,age, cCre,foreign,duration)
#df

#unselcted parameters
#credit_amount <- df$Credit.Amount
#cAdd <-as.factor(df$Duration.in.Current.address)
```

GLMM model 1 - with all available parameters

```
set.seed(221010071)
library(dplyr)
dt = sort(sample(nrow(df), nrow(df)*.7))
df$y <- as.factor(df$y)
#df

#group split
levels(df$y) <- c("Not_Default", "Default")
train<-df[dt,]
test<-df[-dt,]

table(train$y)
```

```
##
## Not_Default     Default
##         489         211
```

```
table(test$y)
```

```
##
## Not_Default     Default
##         211          89
```

```
prop.table(table(train$y))
```

```
##
## Not_Default     Default
##   0.6985714   0.3014286
```

```
prop.table(table(test$y))
```

```
##
## Not_Default     Default
##   0.7033333   0.2966667
```

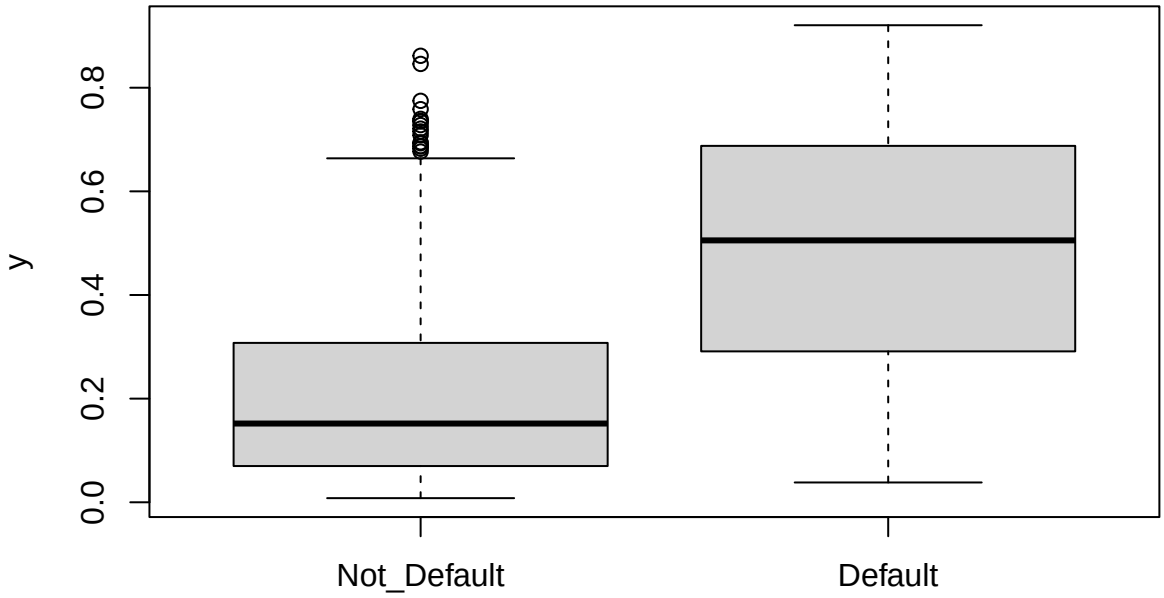
```
#model 1 - everything
model <- lme4::glmer(as.factor(y) ~ accbl + pmt + value + instpct + marrital +mValue + age + cCre  + mV
```

```
## boundary (singular) fit: see help('isSingular')
```

```
# summary(model)
```

```
# model projection on training group
result_train <- predict(model, newdata = train, type = 'response')
trainresult <- data.frame(result_train)
#trainresult
```

```
#boxplot(result_train, df = trainresult)

train_c <- data.frame(train$y, trainresult$result_train)
#train_c

# box plot of prediction on actual
x <- train_c$train.y
y <- train_c$trainresult.result_train
boxplot(y~x)
```




```
# choose proper threshold for default group
library(dplyr)
default <- train_c %>% filter( train_c$train.y == 'Default')
n_default <- train_c %>% filter( train_c$train.y == 'Not_Default')

#d_ll <- median(default$trainresult.result1)
d_ll <- quantile(default$trainresult.result_train, .25)
nd_ul <- median(n_default$trainresult.result_train)

#d_ll
#nd_ul
```

```
# model projection on test group
result_test <- predict(model, newdata = test, type = 'response')
testresult <- data.frame(result_test)
#boxplot(result_test, df = testresult)

# default determination projection on test group
test_c <- data.frame(test$y, testresult$result_test)
test_c$predict_y <- ifelse(result_test >= d_ll,'Default','Not_Default')
test_c$pred_power <- test_c$test.y == test_c$predict_y

#view power of the model
```

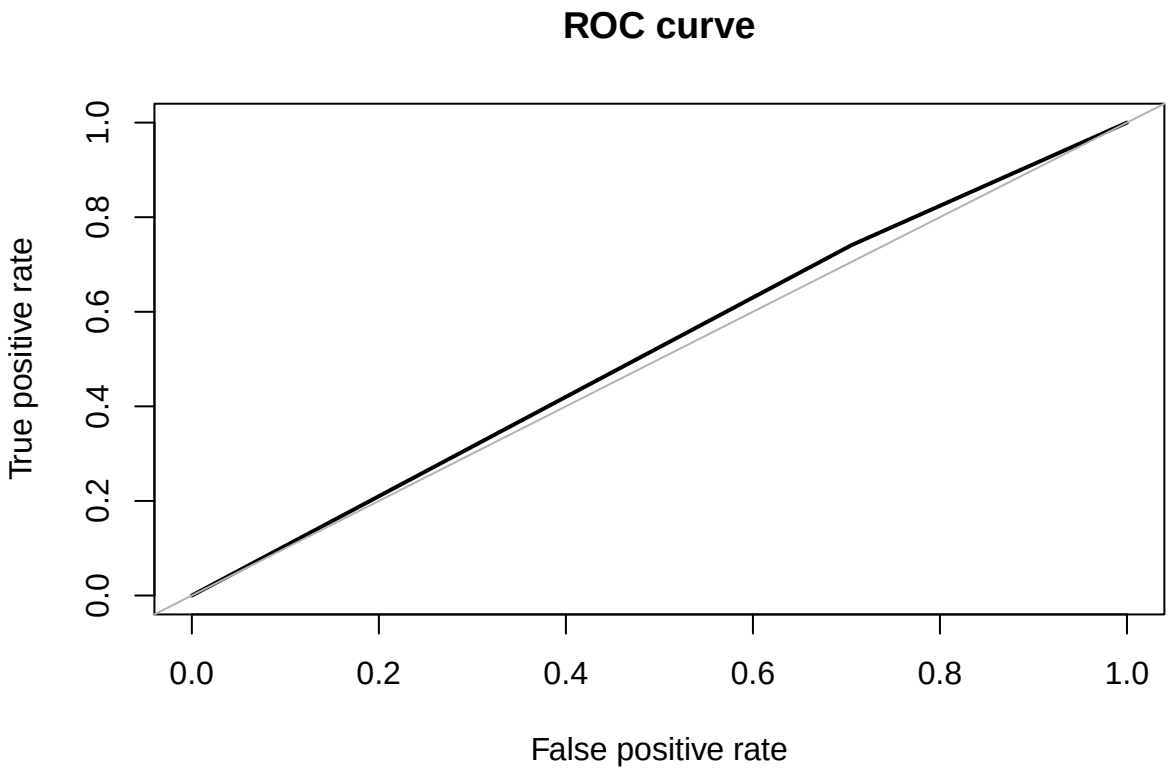
```
table(test_c$test.y,test_c$pred_power)
```

```
##
##                FALSE TRUE
##   Not_Default     62  149
##   Default         23   66
```

```
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

```
roc.curve(test_c$test.y,test_c$pred_power,plotit = TRUE)
```




```
## Area under the curve (AUC): 0.518
```

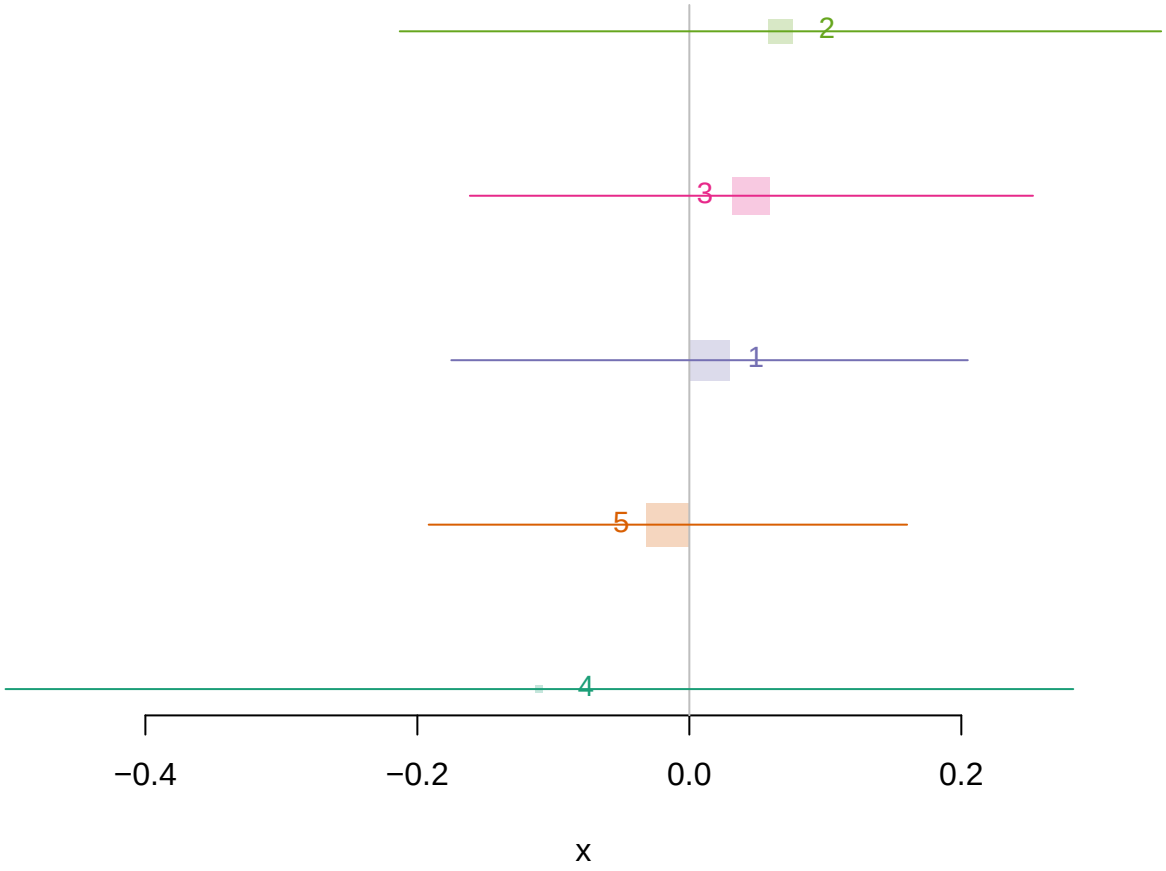
The Area Under the ROC curve (AUC) is an aggregated metric that evaluates how well a logistic regression model classifies positive and negative outcomes at all possible cutoffs. It can range from 0.5 to 1, and the larger it is the better.

```
GLMM1 <- glmmTMB(as.factor(y) ~ accbl + pmt + value + instpct + marrital +mValue + age + cCre  + mValue
summary(GLMM1)
```

```
##  Family: binomial  ( logit )
## Formula:
## as.factor(y) ~ accbl + pmt + value + instpct + marrital + mValue +
##     age + cCre + mValue + duration + (1 | cEmp)
## Data: df
##
##      AIC      BIC   logLik deviance df.resid
##    994.8   1122.4   -471.4    942.8      974
##
## Random effects:
```

```
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  cEmp   (Intercept) 0.01359  0.1166
## Number of obs: 1000, groups:  cEmp, 5
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.059603   0.734469   0.081 0.935322
## accbl2      -0.330417   0.203585  -1.623 0.104591
## accbl3      -0.947227   0.358341  -2.643 0.008209 **
## accbl4      -1.626021   0.221297  -7.348 2.02e-13 ***
## pmt1        -0.184296   0.514927  -0.358 0.720412
## pmt2        -0.848316   0.392678  -2.160 0.030747 *
## pmt3        -0.942684   0.456383  -2.066 0.038871 *
## pmt4        -1.438663   0.418312  -3.439 0.000583 ***
## value2      -0.187604   0.272969  -0.687 0.491911
## value3      -0.480922   0.392328  -1.226 0.220268
## value4      -1.049455   0.493535  -2.126 0.033470 *
## value5      -0.897974   0.246290  -3.646 0.000266 ***
## instpct2     0.120441   0.293532   0.410 0.681574
## instpct3     0.322824   0.316683   1.019 0.308017
## instpct4     0.575158   0.265477   2.167 0.030273 *
## marrital2   -0.187394   0.378877  -0.495 0.620880
## marrital3   -0.736185   0.374185  -1.967 0.049133 *
## marrital4   -0.539806   0.441286  -1.223 0.221233
## mValue2      0.223604   0.238182   0.939 0.347837
## mValue3      0.167116   0.221727   0.754 0.451030
## mValue4      0.669521   0.276845   2.418 0.015589 *
## age         -0.012198   0.008182  -1.491 0.136018
## cCre2       -0.017513   0.406282  -0.043 0.965618
## cCre3       -0.315261   0.233742  -1.349 0.177415
## duration     0.565348   0.081501   6.937 4.01e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Pmisc::ranefPlot(GLMM1, grpvar = "cEmp", level = 0.9, maxNames = 12)
```

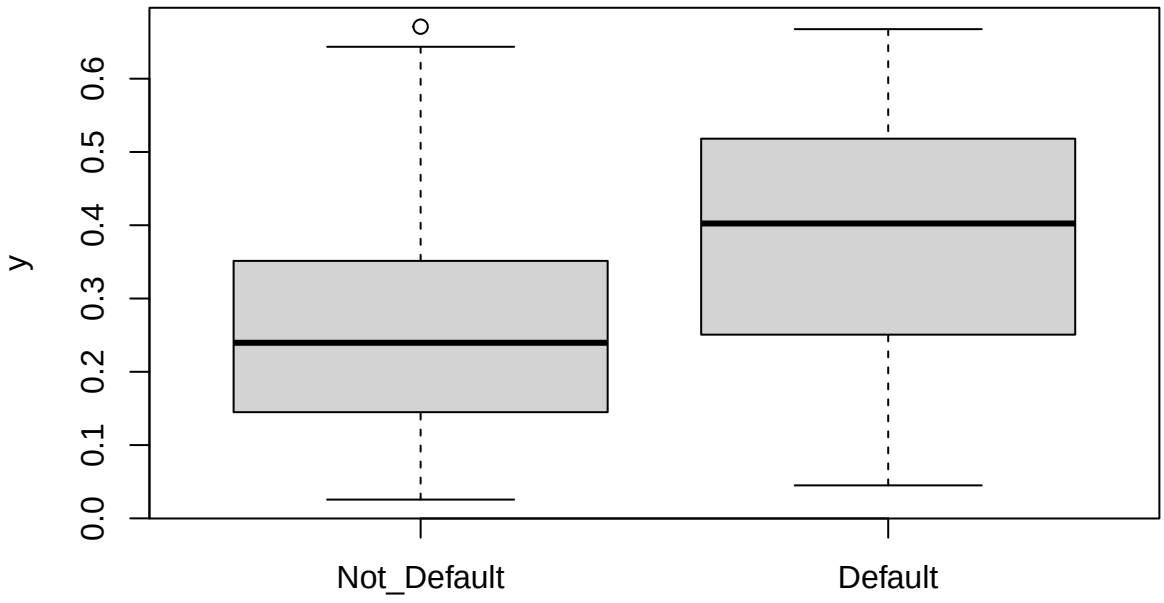


#

GLMM model 2 - with selected parameters

```
#model 2 - selected
model <- lme4::glmer(as.factor(y) ~ age  + value + duration + (1 |cEmp) ,family = binomial, data = trai
```

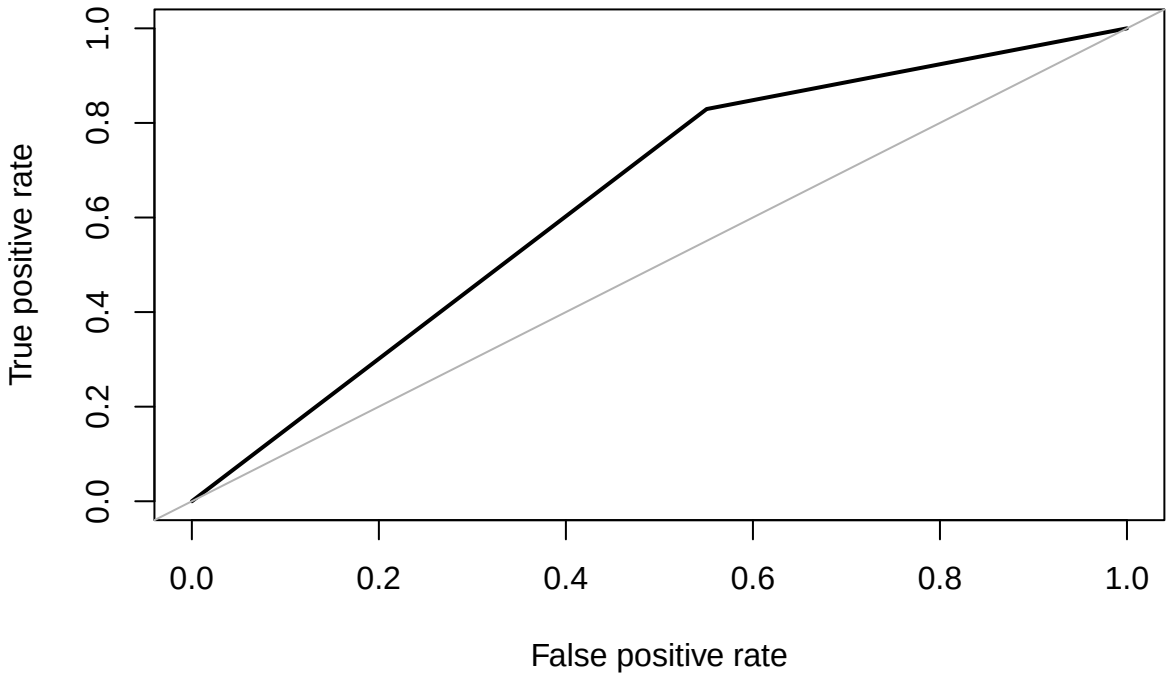
```
result_train <- predict(model, newdata = train, type = 'response')
trainresult <- data.frame(result_train)

train_c <- data.frame(train$y, trainresult$result_train)
x <- train_c$train.y
y <- train_c$trainresult.result_train
boxplot(y~x)
```




```
default <- train_c %>% filter( train_c$train.y == 'Default')
n_default <- train_c %>% filter( train_c$train.y == 'Not_Default')

#d_ll <- median(default$trainresult.result1)
d_ll <- quantile(default$trainresult.result_train, .5)
nd_ul <- median(n_default$trainresult.result_train)

# model projection on test group
result_test <- predict(model, newdata = test, type = 'response')
testresult <- data.frame(result_test)
#boxplot(result_test, df = testresult)

# default determination projection on test group
test_c <- data.frame(test$y, testresult$result_test)
test_c$predict_y <- ifelse(result_test >= d_ll,'Default','Not_Default')
test_c$pred_power <- test_c$test.y == test_c$predict_y

#view power of the model
table(test_c$test.y,test_c$pred_power)
```

```
##
##                FALSE TRUE
##   Not_Default     36  175
##   Default         40   49
```

```
roc.curve(test_c$test.y,test_c$pred_power,plotit = TRUE)
```

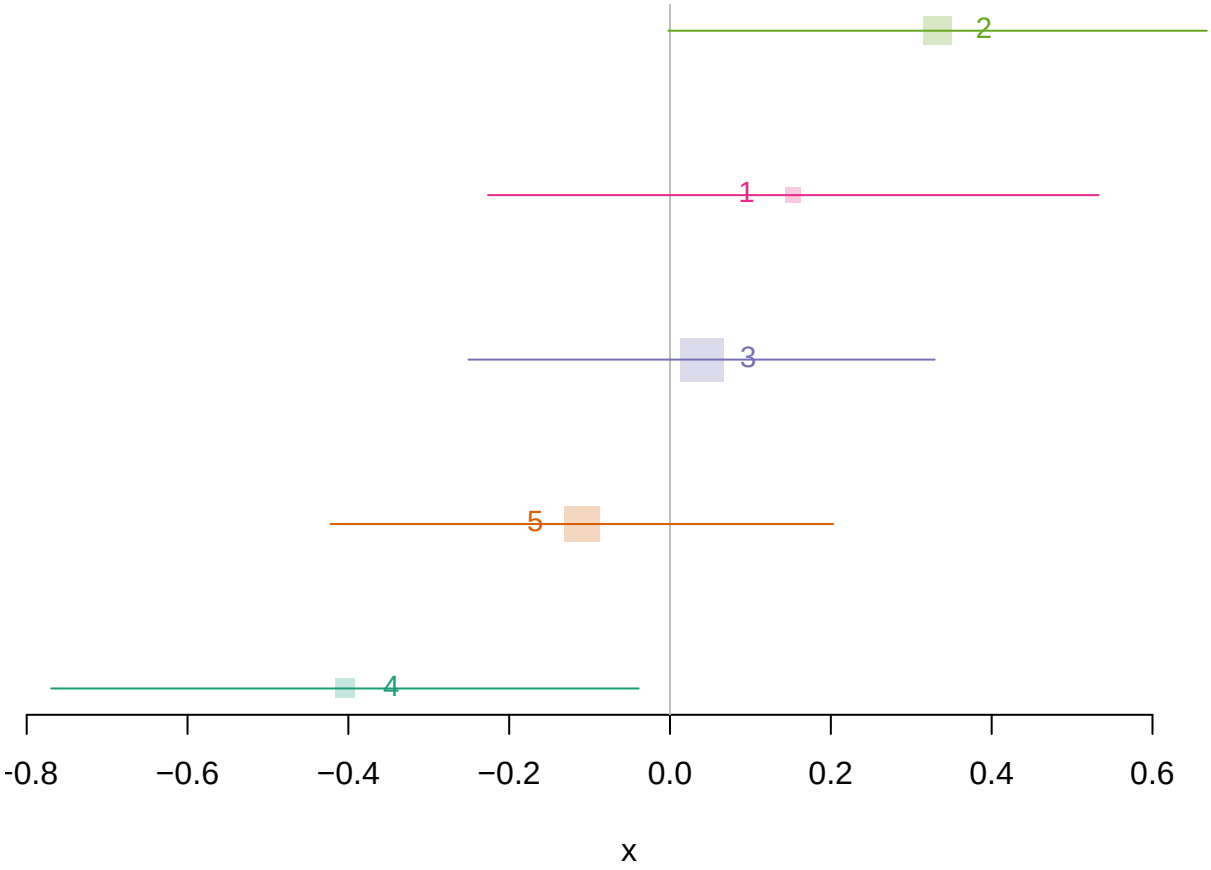



```
## Area under the curve (AUC): 0.639
```

```
GLMM2 <- glmmTMB(as.factor(y) ~ age  + value + duration + (1 |cEmp) , data = df, family = binomial(link
summary(GLMM2)
```

```
##  Family: binomial  ( logit )
## Formula:          as.factor(y) ~ age + value + duration + (1 | cEmp)
## Data: df
##
##      AIC      BIC   logLik deviance df.resid
##   1097.4   1136.6   -540.7   1081.4      992
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  cEmp   (Intercept) 0.08671  0.2945
## Number of obs: 1000, groups:  cEmp, 5
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.890253   0.369286  -5.119 3.08e-07 ***
## age         -0.011077   0.007463  -1.484  0.13772
## value2      -0.184140   0.241229  -0.763  0.44526
## value3      -0.906930   0.359282  -2.524  0.01159 *
## value4      -1.276615   0.458456  -2.785  0.00536 **
## value5      -1.101582   0.223483  -4.929 8.26e-07 ***
## duration     0.653017   0.071629   9.117  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
install.packages("Pmisc", repos = "http://R-Forge.R-project.org", type = "source")
Pmisc::ranefPlot(GLMM2, grpvar = "cEmp", level = 0.90, maxNames = 12)
```



Decision tree selection (model arrange in accending AUC)

Decision tree model 1 with all available parameters

```
library(C50)
model <- C5.0(train$y~. ,data = train)
summary(model)
```

```
##
## Call:
## C5.0.formula(formula = train$y ~ ., data = train)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Apr 26 23:47:40 2022
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 700 cases (12 attributes) from undefined.data
##
## Decision tree:
##
```

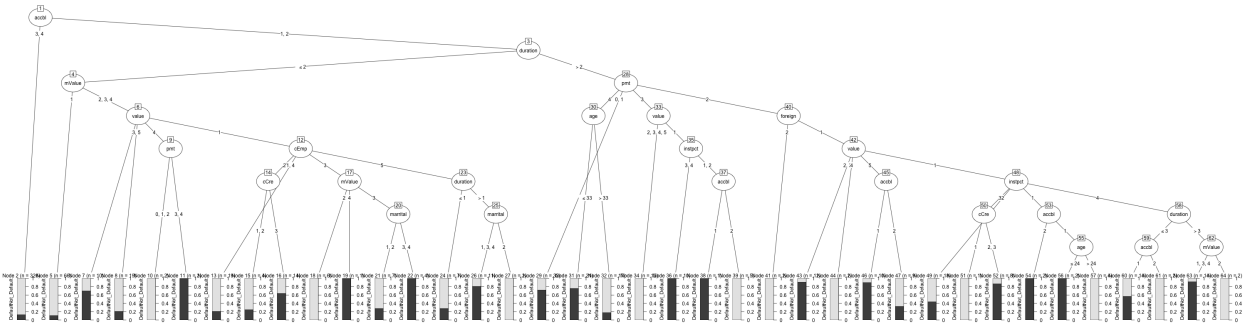
```
## accbl in {3,4}: Not_Default (326/43)
## accbl in {1,2}:
## :...duration <= 2:
##     :...mValue = 1: Not_Default (66/7)
##     :   mValue in {2,3,4}:
##     :   :...value = 2: Default (10/3)
##     :       value in {3,5}: Not_Default (19/4)
##     :       value = 4:
##     :       :...pmt in {0,1,2}: Not_Default (2)
##     :       :   pmt in {3,4}: Default (2)
##     :       value = 1:
##     :       :...cEmp in {1,4}: Not_Default (19/4)
##     :           cEmp = 2:
##     :           :...cCre in {1,2}: Not_Default (4/1)
##     :           :   cCre = 3: Default (14/5)
##     :           cEmp = 3:
##     :           :...mValue = 2: Not_Default (6)
##     :           :   mValue = 4: Default (1)
##     :           :   mValue = 3:
##     :           :   :...marrital in {1,2}: Not_Default (7/2)
##     :           :       marrital in {3,4}: Default (4)
##     :           cEmp = 5:
##     :           :...duration <= 1: Not_Default (7/2)
##     :               duration > 1:
##     :               :...marrital in {1,3,4}: Default (11/2)
##     :                   marrital = 2: Not_Default (2)
##     duration > 2:
##     :...pmt in {0,1}: Default (32/9)
##         pmt = 4:
##         :...age <= 33: Default (21/5)
##         :   age > 33: Not_Default (17/3)
##         pmt = 3:
##         :...value in {2,3,4,5}: Not_Default (12)
##         :   value = 1:
##         :   :...instpct in {3,4}: Default (10)
##         :       instpct in {1,2}:
##         :       :...accbl = 1: Default (1)
##         :           accbl = 2: Not_Default (5)
##         pmt = 2:
##         :...foreign = 2: Not_Default (2)
##             foreign = 1:
##             :...value in {2,3}: Default (12/1)
##                 value = 4: Not_Default (2)
##                 value = 5:
##                 :...accbl = 1: Default (10/1)
##                 :   accbl = 2: Not_Default (9/3)
##                 value = 1:
##                 :...instpct = 2: Not_Default (18/8)
##                     instpct = 3:
##                     :...cCre = 1: Not_Default (1)
##                     :   cCre in {2,3}: Default (8/1)
##                     instpct = 1:
##                     :...accbl = 2: Default (2)
##                     :   accbl = 1:
```

```
##                     :   :...age <= 24: Default (2)
##                     :       age > 24: Not_Default (4)
##                     instpct = 4:
##                     :...duration <= 3:
##                         :...accbl = 1: Default (14/6)
##                         :   accbl = 2: Not_Default (2)
##                         duration > 3:
##                         :...mValue in {1,3,4}: Default (14/1)
##                             mValue = 2: Not_Default (2)
##
##
## Evaluation on training data (700 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      38  111(15.9%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     455    34    (a): class Not_Default
##      77   134    (b): class Default
##
##
##  Attribute usage:
##
##  100.00% accbl
##   53.43% duration
##   33.71% value
##   29.14% pmt
##   27.14% mValue
##   14.57% foreign
##   11.86% instpct
##   10.71% cEmp
##    6.29% age
##    3.86% cCre
##    3.43% marrital
##
##
## Time: 0.0 secs
```

```
#train

png("decision_tree1.png", width = 3000, height = 800)
plot(model)
dev.off()
```

```
## pdf
##   2
```

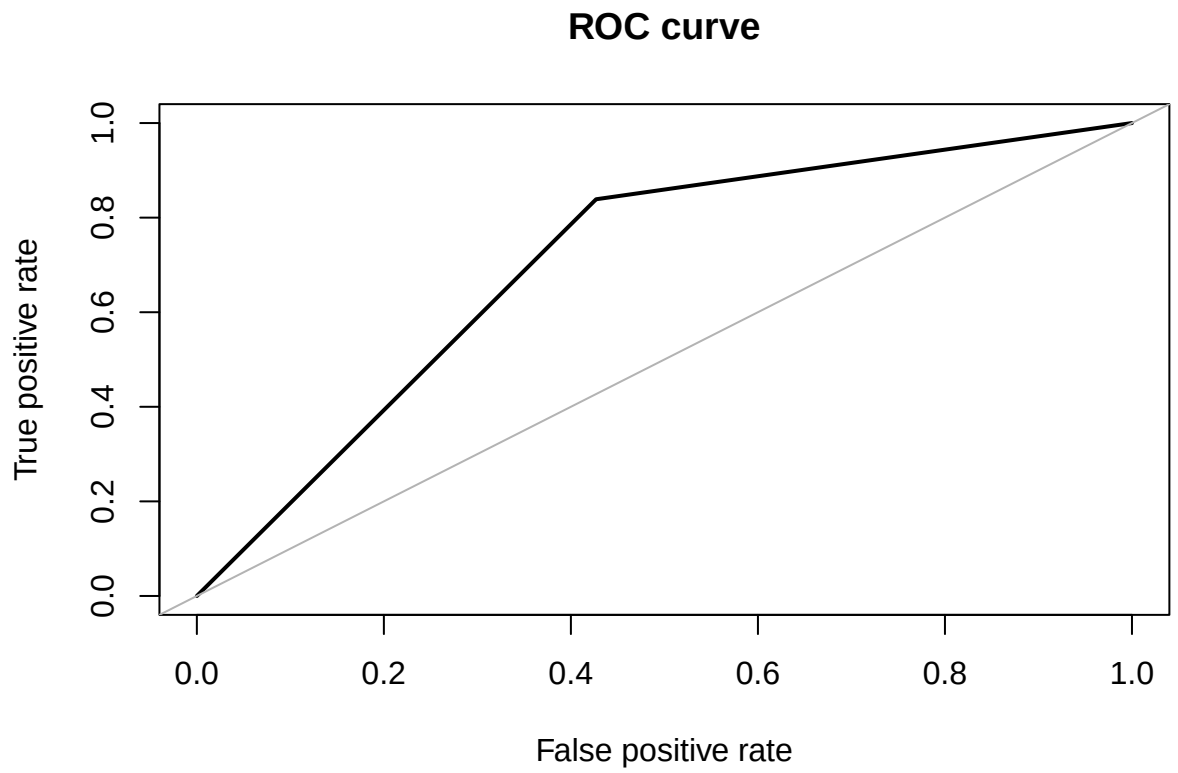
```
knitr::include_graphics("decision_tree1.png")
```



```
result_test <- predict(model, newdata = test, trails = 100, type = 'class')
testresult <- data.frame(result_test)
test_c <- data.frame(test$y, testresult$result_test)
test_c$pred_power <- test_c$test.y == test_c$testresult.result_test
table(test_c$test.y,test_c$pred_power)
```

```
##
##                FALSE TRUE
##   Not_Default     34  177
##   Default         51   38
```

```
library(ROSE)
roc.curve(test_c$test.y,test_c$pred_power,plotit = TRUE)
```



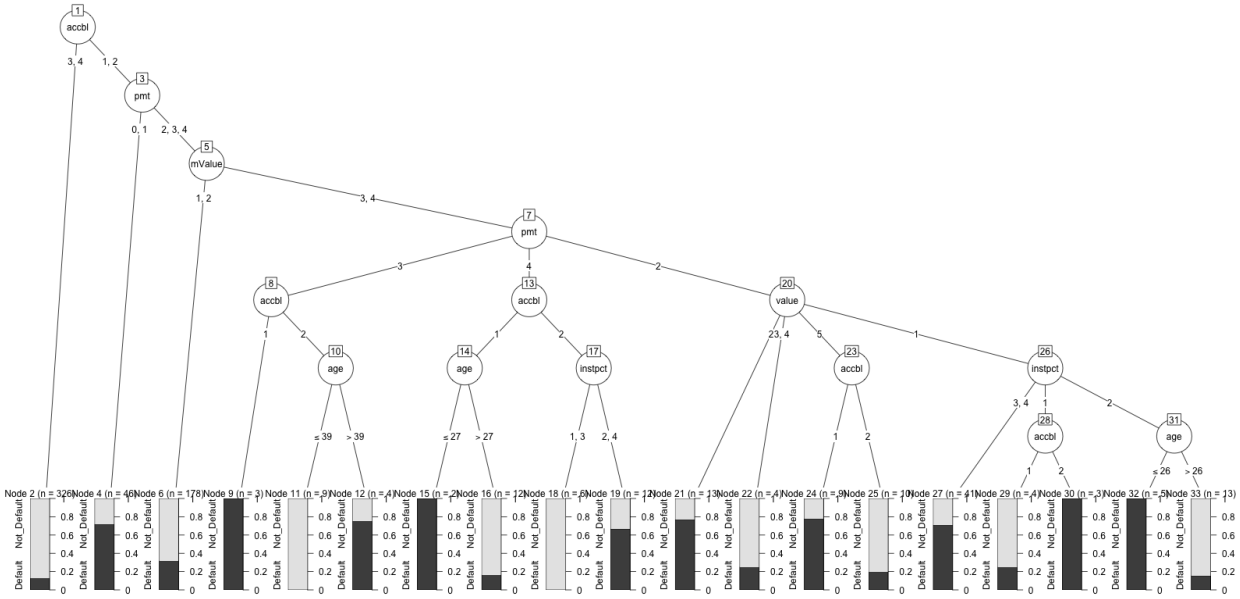

```
## Area under the curve (AUC): 0.706
```

Decision tree model 2 with selected variables(from model 1)

```
model2 <- C5.0(train$y~ value+pmt+cCre+mValue+marrital+age+instpct+accbl ,data = train)
#summary(model)

png("decision_tree2.png", width = 1600, height = 800)
plot(model2)
dev.off()
```

```
## pdf
##   2
```

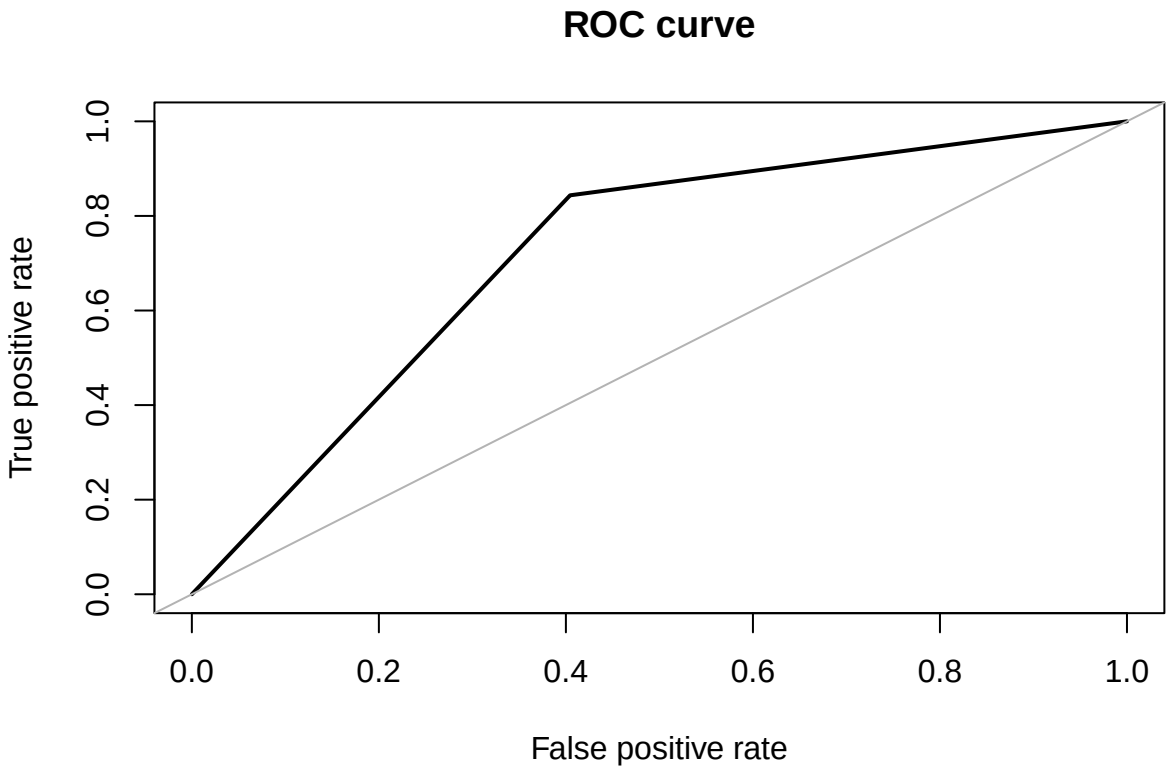
```
knitr::include_graphics("decision_tree2.png")
```




```
result_test <- predict(model2, newdata = test, trails = 100, type = 'class')
testresult <- data.frame(result_test)
test_c <- data.frame(test$y, testresult$result_test)
test_c$pred_power <- test_c$test.y == test_c$testresult.result_test
table(test_c$test.y,test_c$pred_power)
```

```
##
##                FALSE TRUE
##   Not_Default     33  178
##   Default         53   36
```

```
library(ROSE)
roc.curve(test_c$test.y,test_c$pred_power,plotit = TRUE)
```



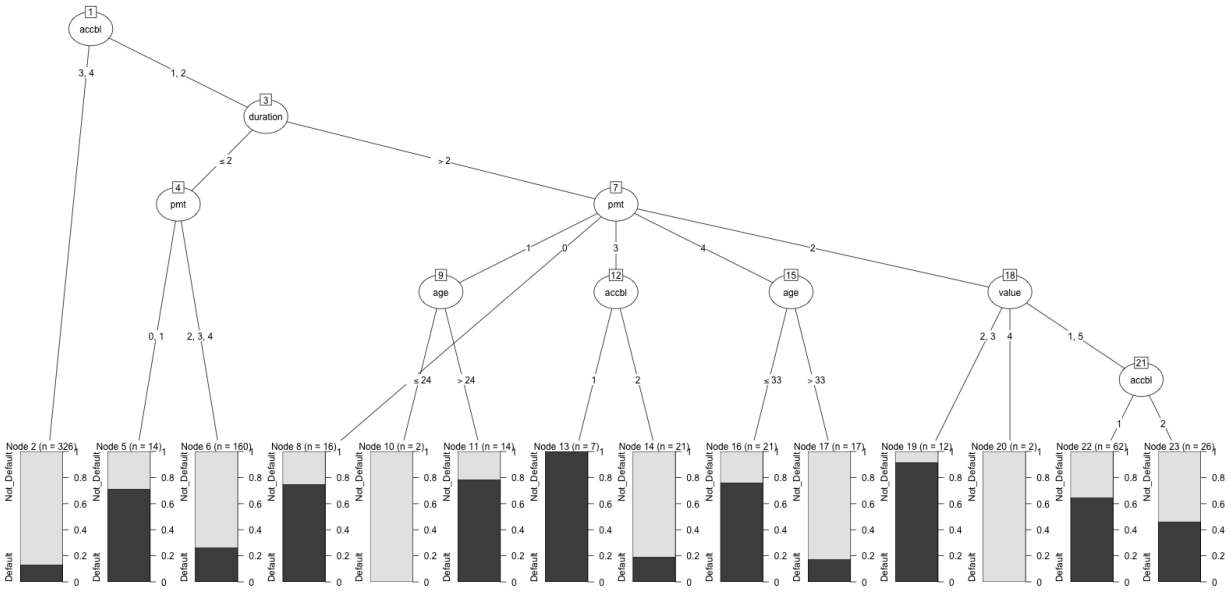

```
## Area under the curve (AUC): 0.720
```

Decision tree model 3

```
model3 <- C5.0(train$y~ accbl + duration + value + pmt + age ,data = train)
#summary(model)

png("decision_tree3.png", width = 1600, height = 800)
plot(model3)
dev.off()
```

```
## pdf
##   2
```

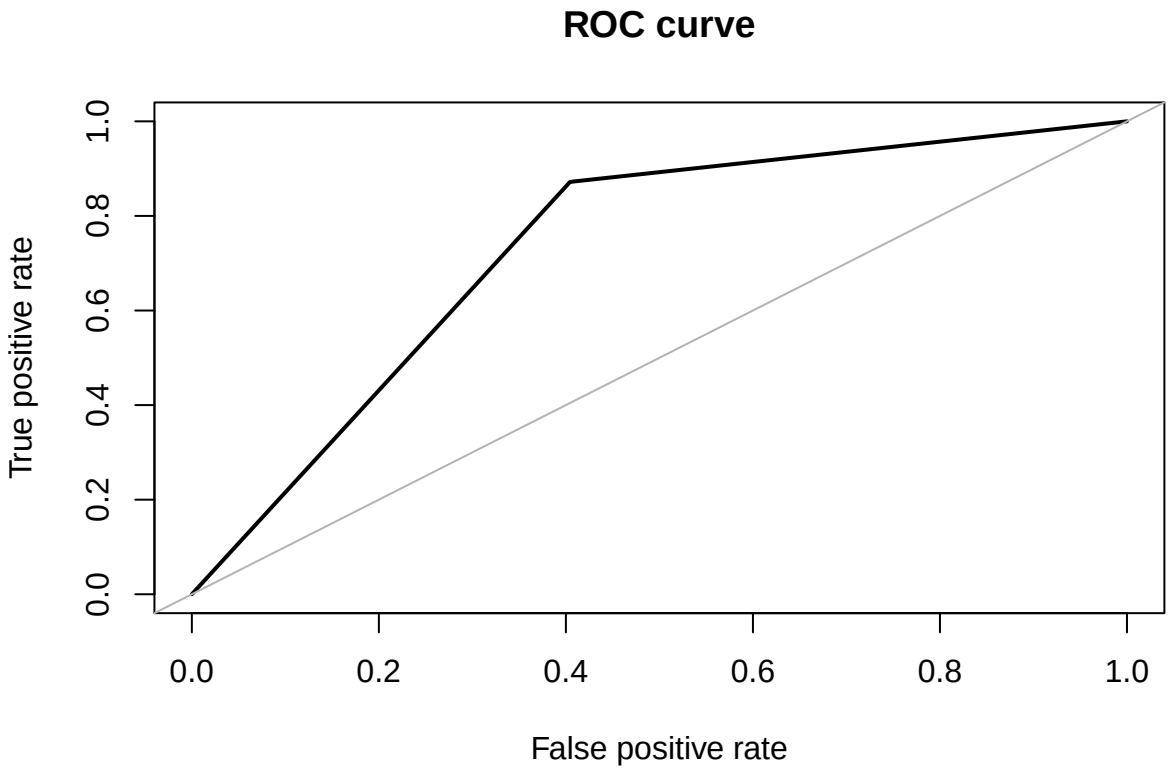
```
knitr::include_graphics("decision_tree3.png")
```




```
result_test <- predict(model3, newdata = test, trails = 100, type = 'class')
testresult <- data.frame(result_test)
test_c <- data.frame(test$y, testresult$result_test)
test_c$pred_power <- test_c$test.y == test_c$testresult.result_test
table(test_c$test.y,test_c$pred_power)
```

```
##
##                FALSE TRUE
##   Not_Default     27  184
##   Default         53   36
```

```
library(ROSE)
roc.curve(test_c$test.y,test_c$pred_power,plotit = TRUE)
```




```
## Area under the curve (AUC): 0.734
```

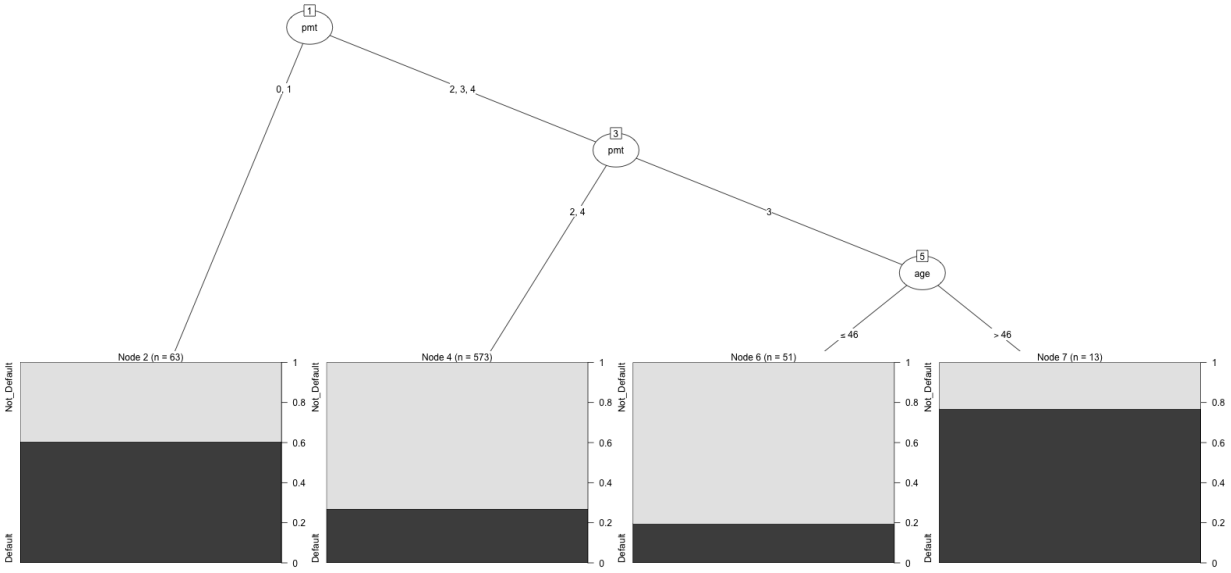
Decision tree model 4 with selected variables(optimal)

Based on model 1, subtrees in marital, value and age have good performance, and therefore they are obtained for further modelling.

```
model4 <- C5.0(train$y~ marrital +value+age +pmt ,data = train)
#summary(model)

png("decision_tree4.png", width = 1600, height = 800)
plot(model4)
dev.off()
```

```
## pdf
##   2
```

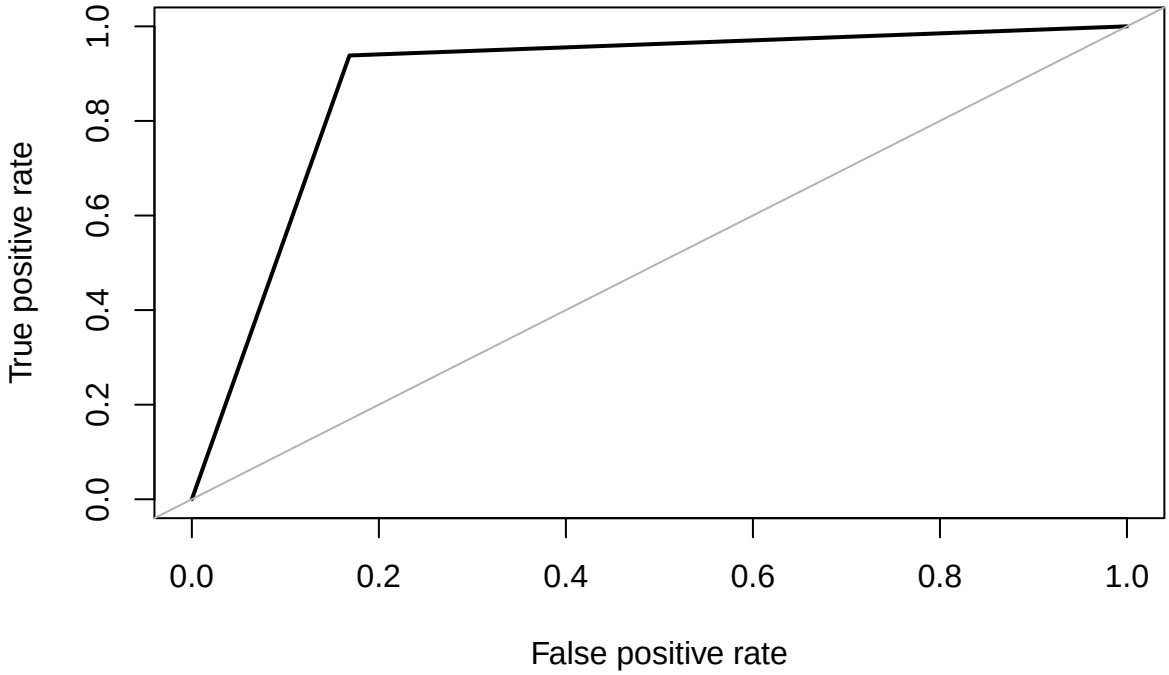
```
knitr::include_graphics("decision_tree4.png")
```




```
result_test <- predict(model4, newdata = test, trails = 100, type = 'class')
testresult <- data.frame(result_test)
test_c <- data.frame(test$y, testresult$result_test)
test_c$pred_power <- test_c$test.y == test_c$testresult.result_test
table(test_c$test.y,test_c$pred_power)
```

```
##
##                FALSE TRUE
##   Not_Default     13  198
##   Default         74   15
```

```
roc.curve(test_c$test.y,test_c$pred_power,plotit = TRUE)
```




```
## Area under the curve (AUC): 0.885
```