Classification Tree, Random Forest

# Download the data

set.seed(123)  
setwd('C:/Users/VerkhovodTS/Desktop/clas')  
f <- read.csv2('clients.csv', header = TRUE, encoding = 'UNICOD')

Висновок: використані дані про наявність прострочених платежів по кредиту.

# Statistics

## Descriptive statistics

library (psych)

## Warning: package 'psych' was built under R version 3.6.3

describe(f)

## vars n mean sd median trimmed  
## LOAN\_AMOUNT 1 4000 21470.59 20885.55 16816.14 17493.79  
## BRANCH\_REGION\* 2 4000 11.22 4.97 12.00 11.48  
## CLIENT\_GENDER\* 3 4000 1.49 0.50 1.00 1.49  
## CLIENT\_FAMILYSTATUS\* 4 4000 4.54 1.11 5.00 4.68  
## CLIENT\_TOTALEXPERIENCE 5 4000 18.93 12.30 17.00 18.24  
## CLIENT\_EDUCATION\* 6 4000 3.75 2.31 5.00 3.81  
## CLIENT\_ACTIVITYTYPE\* 7 4000 4.32 1.59 4.00 4.42  
## CLIENT\_TOGETHER.INCOME 8 4000 10062.58 14185.66 7000.00 7404.52  
## LOAN\_OUTSTANDINGLOANSCOUNT 9 4000 3.01 2.48 2.00 2.67  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 10 4000 1.69 0.46 2.00 1.73  
## EMPLOYMENTTYPE\* 11 4000 2.22 0.84 2.00 2.00  
## LOAN.TERM 12 4000 794.45 341.65 729.00 803.05  
## ZODIAC\* 13 4000 6.25 3.44 6.00 6.21  
## ZODIAC\_CHINA\* 14 4000 6.62 3.43 7.00 6.65  
## AGE 15 4000 40.78 13.27 40.00 40.25  
## CHANGE.WORK\* 16 4000 1.16 0.36 1.00 1.07  
## REAL\_ESTATE\* 17 4000 1.39 0.49 1.00 1.36  
## CAR\* 18 4000 1.91 0.29 2.00 2.00  
## DELAY\* 19 4000 1.58 0.49 2.00 1.60  
## mad min max range skew kurtosis  
## LOAN\_AMOUNT 14958.96 1000.0 116822.4 115822.4 2.75 8.55  
## BRANCH\_REGION\* 4.45 1.0 19.0 18.0 -0.46 -0.77  
## CLIENT\_GENDER\* 0.00 1.0 2.0 1.0 0.03 -2.00  
## CLIENT\_FAMILYSTATUS\* 1.48 1.0 6.0 5.0 -1.37 2.15  
## CLIENT\_TOTALEXPERIENCE 14.83 0.1 52.0 51.9 0.38 -0.97  
## CLIENT\_EDUCATION\* 1.48 1.0 6.0 5.0 -0.25 -1.82  
## CLIENT\_ACTIVITYTYPE\* 2.97 1.0 7.0 6.0 -0.30 -1.14  
## CLIENT\_TOGETHER.INCOME 3203.57 1800.0 249464.5 247664.5 7.04 71.88  
## LOAN\_OUTSTANDINGLOANSCOUNT 1.48 0.0 21.0 21.0 1.85 5.89  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 0.00 1.0 2.0 1.0 -0.80 -1.36  
## EMPLOYMENTTYPE\* 0.00 1.0 5.0 4.0 2.72 6.48  
## LOAN.TERM 541.15 60.0 2007.0 1947.0 -0.16 -1.12  
## ZODIAC\* 4.45 1.0 12.0 11.0 0.08 -1.18  
## ZODIAC\_CHINA\* 4.45 1.0 12.0 11.0 -0.06 -1.20  
## AGE 16.31 21.0 69.0 48.0 0.25 -1.06  
## CHANGE.WORK\* 0.00 1.0 2.0 1.0 1.89 1.57  
## REAL\_ESTATE\* 0.00 1.0 2.0 1.0 0.47 -1.78  
## CAR\* 0.00 1.0 2.0 1.0 -2.87 6.24  
## DELAY\* 0.00 1.0 2.0 1.0 -0.31 -1.91  
## se  
## LOAN\_AMOUNT 330.23  
## BRANCH\_REGION\* 0.08  
## CLIENT\_GENDER\* 0.01  
## CLIENT\_FAMILYSTATUS\* 0.02  
## CLIENT\_TOTALEXPERIENCE 0.19  
## CLIENT\_EDUCATION\* 0.04  
## CLIENT\_ACTIVITYTYPE\* 0.03  
## CLIENT\_TOGETHER.INCOME 224.29  
## LOAN\_OUTSTANDINGLOANSCOUNT 0.04  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 0.01  
## EMPLOYMENTTYPE\* 0.01  
## LOAN.TERM 5.40  
## ZODIAC\* 0.05  
## ZODIAC\_CHINA\* 0.05  
## AGE 0.21  
## CHANGE.WORK\* 0.01  
## REAL\_ESTATE\* 0.01  
## CAR\* 0.00  
## DELAY\* 0.01

Висновок: кількість спостережень – 4000, кількість змінних – 19, з них якісних – 13, кількісних – 6. Пропущених значень немає. Викиди мають змінні: LOAN\_AMOUNT, CLIENT\_TOGETHER.INCOME, LOAN\_OUTSTANDINGLOANSCOUNT

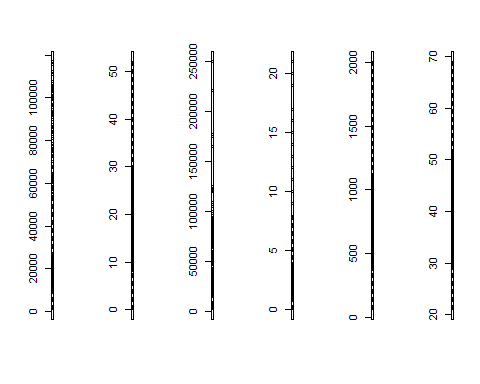
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

par(mfrow = c(1,6))  
boxplot(f$LOAN\_AMOUNT)  
boxplot(f$CLIENT\_TOTALEXPERIENCE)  
boxplot(f$CLIENT\_TOGETHER.INCOME)  
boxplot(f$LOAN\_OUTSTANDINGLOANSCOUNT)  
boxplot(f$LOAN.TERM)  
boxplot(f$AGE)



Висновок: Викиди мають змінні: LOAN\_AMOUNT, CLIENT\_TOGETHER.INCOME, LOAN\_OUTSTANDINGLOANSCOUNT

## Ejections (outside the three sigma)

f\_ej <- f  
f\_ej$LOAN\_AMOUNT<- ifelse(f$LOAN\_AMOUNT < mean(f$LOAN\_AMOUNT)+sd(f$LOAN\_AMOUNT)\*3, f$LOAN\_AMOUNT, mean(f$LOAN\_AMOUNT)+sd(f$LOAN\_AMOUNT)\*3)  
f\_ej$CLIENT\_TOGETHER.INCOME<- ifelse(f$CLIENT\_TOGETHER.INCOME < mean(f$CLIENT\_TOGETHER.INCOME)+sd(f$CLIENT\_TOGETHER.INCOME)\*3, f$CLIENT\_TOGETHER.INCOME, mean(f$CLIENT\_TOGETHER.INCOME)+sd(f$CLIENT\_TOGETHER.INCOME)\*3)  
f\_ej$LOAN\_OUTSTANDINGLOANSCOUNT<- ifelse(f$LOAN\_OUTSTANDINGLOANSCOUNT < mean(f$LOAN\_OUTSTANDINGLOANSCOUNT)+sd(f$LOAN\_OUTSTANDINGLOANSCOUNT)\*3, f$LOAN\_OUTSTANDINGLOANSCOUNT, mean(f$LOAN\_OUTSTANDINGLOANSCOUNT)+sd(f$LOAN\_OUTSTANDINGLOANSCOUNT)\*3)  
describe(f\_ej)

## vars n mean sd median trimmed  
## LOAN\_AMOUNT 1 4000 20653.71 17666.41 16816.14 17493.79  
## BRANCH\_REGION\* 2 4000 11.22 4.97 12.00 11.48  
## CLIENT\_GENDER\* 3 4000 1.49 0.50 1.00 1.49  
## CLIENT\_FAMILYSTATUS\* 4 4000 4.54 1.11 5.00 4.68  
## CLIENT\_TOTALEXPERIENCE 5 4000 18.93 12.30 17.00 18.24  
## CLIENT\_EDUCATION\* 6 4000 3.75 2.31 5.00 3.81  
## CLIENT\_ACTIVITYTYPE\* 7 4000 4.32 1.59 4.00 4.42  
## CLIENT\_TOGETHER.INCOME 8 4000 9288.06 8925.54 7000.00 7404.52  
## LOAN\_OUTSTANDINGLOANSCOUNT 9 4000 2.96 2.28 2.00 2.67  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 10 4000 1.69 0.46 2.00 1.73  
## EMPLOYMENTTYPE\* 11 4000 2.22 0.84 2.00 2.00  
## LOAN.TERM 12 4000 794.45 341.65 729.00 803.05  
## ZODIAC\* 13 4000 6.25 3.44 6.00 6.21  
## ZODIAC\_CHINA\* 14 4000 6.62 3.43 7.00 6.65  
## AGE 15 4000 40.78 13.27 40.00 40.25  
## CHANGE.WORK\* 16 4000 1.16 0.36 1.00 1.07  
## REAL\_ESTATE\* 17 4000 1.39 0.49 1.00 1.36  
## CAR\* 18 4000 1.91 0.29 2.00 2.00  
## DELAY\* 19 4000 1.58 0.49 2.00 1.60  
## mad min max range skew kurtosis  
## LOAN\_AMOUNT 14958.96 1000.0 84127.24 83127.24 2.12 4.96  
## BRANCH\_REGION\* 4.45 1.0 19.00 18.00 -0.46 -0.77  
## CLIENT\_GENDER\* 0.00 1.0 2.00 1.00 0.03 -2.00  
## CLIENT\_FAMILYSTATUS\* 1.48 1.0 6.00 5.00 -1.37 2.15  
## CLIENT\_TOTALEXPERIENCE 14.83 0.1 52.00 51.90 0.38 -0.97  
## CLIENT\_EDUCATION\* 1.48 1.0 6.00 5.00 -0.25 -1.82  
## CLIENT\_ACTIVITYTYPE\* 2.97 1.0 7.00 6.00 -0.30 -1.14  
## CLIENT\_TOGETHER.INCOME 3203.57 1800.0 52619.55 50819.55 3.53 13.42  
## LOAN\_OUTSTANDINGLOANSCOUNT 1.48 0.0 10.44 10.44 1.19 1.32  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 0.00 1.0 2.00 1.00 -0.80 -1.36  
## EMPLOYMENTTYPE\* 0.00 1.0 5.00 4.00 2.72 6.48  
## LOAN.TERM 541.15 60.0 2007.00 1947.00 -0.16 -1.12  
## ZODIAC\* 4.45 1.0 12.00 11.00 0.08 -1.18  
## ZODIAC\_CHINA\* 4.45 1.0 12.00 11.00 -0.06 -1.20  
## AGE 16.31 21.0 69.00 48.00 0.25 -1.06  
## CHANGE.WORK\* 0.00 1.0 2.00 1.00 1.89 1.57  
## REAL\_ESTATE\* 0.00 1.0 2.00 1.00 0.47 -1.78  
## CAR\* 0.00 1.0 2.00 1.00 -2.87 6.24  
## DELAY\* 0.00 1.0 2.00 1.00 -0.31 -1.91  
## se  
## LOAN\_AMOUNT 279.33  
## BRANCH\_REGION\* 0.08  
## CLIENT\_GENDER\* 0.01  
## CLIENT\_FAMILYSTATUS\* 0.02  
## CLIENT\_TOTALEXPERIENCE 0.19  
## CLIENT\_EDUCATION\* 0.04  
## CLIENT\_ACTIVITYTYPE\* 0.03  
## CLIENT\_TOGETHER.INCOME 141.13  
## LOAN\_OUTSTANDINGLOANSCOUNT 0.04  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 0.01  
## EMPLOYMENTTYPE\* 0.01  
## LOAN.TERM 5.40  
## ZODIAC\* 0.05  
## ZODIAC\_CHINA\* 0.05  
## AGE 0.21  
## CHANGE.WORK\* 0.01  
## REAL\_ESTATE\* 0.01  
## CAR\* 0.00  
## DELAY\* 0.01

f <- f\_ej

Висновок: для корекції викидів обраний варіант заповнення граничними значеннями.

## Factors as numeric

f$BRANCH\_REGION <- as.numeric(as.factor(f$BRANCH\_REGION))-1  
f$CLIENT\_GENDER <- as.numeric(as.factor(f$CLIENT\_GENDER))-1  
f$CLIENT\_FAMILYSTATUS <- as.numeric(as.factor(f$CLIENT\_FAMILYSTATUS))-1  
f$CLIENT\_EDUCATION <- as.numeric(as.factor(f$CLIENT\_EDUCATION))-1  
f$CLIENT\_ACTIVITYTYPE <- as.numeric(as.factor(f$CLIENT\_ACTIVITYTYPE))-1  
f$LOAN\_OVERDUE\_EXIST\_FLAG <- as.numeric(as.factor(f$LOAN\_OVERDUE\_EXIST\_FLAG))-1  
f$EMPLOYMENTTYPE <- as.numeric(as.factor(f$EMPLOYMENTTYPE))-1  
f$ZODIAC <- as.numeric(as.factor(f$ZODIAC))-1  
f$ZODIAC\_CHINA <- as.numeric(as.factor(f$ZODIAC\_CHINA))-1  
f$CHANGE.WORK <- as.numeric(as.factor(f$CHANGE.WORK))-1  
f$REAL\_ESTATE <- as.numeric(as.factor(f$REAL\_ESTATE))-1  
f$CAR <- as.numeric(as.factor(f$CAR))-1  
f$DELAY <- as.numeric(as.factor(f$DELAY))-1

Висновок: якісні показники були перведені у кількісні. (DELAY=1-немає заборгованості,DELAY=0 - є забаргованість)

# Splitting the scaled dataset into the TRAIN set and TEST set

set.seed(123)  
library(caTools)

## Warning: package 'caTools' was built under R version 3.6.3

split = sample.split(f$DELAY, SplitRatio = 0.8)  
f\_train = subset(f, split == TRUE)  
f\_test = subset(f, split == FALSE)

Висновок: підготований датасет розділено на навчальну та тестову вибірки.

# Fitting

# install.packages('rpart')  
library(rpart)

## Warning: package 'rpart' was built under R version 3.6.3

f\_train$DELAY <- as.factor(f\_train$DELAY)  
f\_test$DELAY <- as.factor(f\_test$DELAY)  
class\_dt = rpart(DELAY ~ ., data = f\_train)

Висновок: для навчання моделі використано функцію rpart.

# Predicting

y <- predict(class\_dt, f\_test[-20], type = 'class')

Висновок: визначені класи об’єктів (вектор у).

# Confusion Matrix

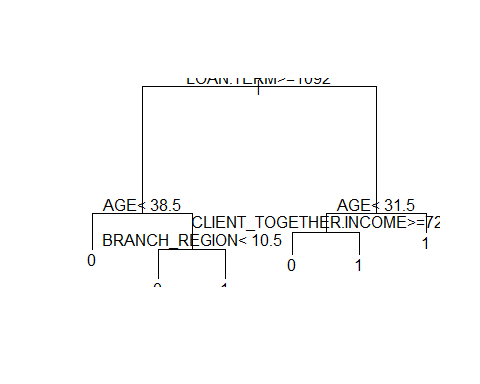
cm4 = table(f\_test[, 'DELAY'], y)  
print(cm4)

## y  
## 0 1  
## 0 204 135  
## 1 120 341

Висновок: точність моделі - 68.13%, частка невірно класифікованих випадків – 31.88%. Чутливість моделі – 73.97%, специфічність – 60.18%, тобто модель більш чутлива до виявлення позитивних випадків (DELAY=1, прострочки немає).

# Plotting the tree

plot(class\_dt)  
text(class\_dt)



Висновок: візуалізація дозволяє проаналізувати логіку побудови дерева.

# Fitting Random Forest Classification to the Training set

# install.packages ('randomForest')  
library(randomForest)

## Warning: package 'randomForest' was built under R version 3.6.3

## randomForest 4.6-14

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

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:psych':  
##   
## outlier

set.seed(123)  
class\_rf = randomForest(DELAY ~ BRANCH\_REGION + CLIENT\_GENDER + CLIENT\_FAMILYSTATUS +   
 CLIENT\_TOTALEXPERIENCE + CLIENT\_EDUCATION + CLIENT\_ACTIVITYTYPE + LOAN\_OUTSTANDINGLOANSCOUNT  
 + LOAN.TERM, data = f\_train, ntree = 50)

Висновок: для навчання моделі використано функцію randomForest.

# Predicting

y <- predict(class\_rf, f\_test[, c('BRANCH\_REGION', 'CLIENT\_GENDER', 'CLIENT\_FAMILYSTATUS', 'CLIENT\_TOTALEXPERIENCE', 'CLIENT\_EDUCATION', 'CLIENT\_ACTIVITYTYPE', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM')])

Висновок: визначені класи об’єктів (вектор у).

# Confusion Matrix

cm8 = table(f\_test[, 'DELAY'], y)  
print(cm8)

## y  
## 0 1  
## 0 211 128  
## 1 93 368

Висновок: точність моделі - 72.38%, частка невірно класифікованих випадків – 27.63%. Чутливість моделі – 79.83%, специфічність – 62.24%, тобто модель більш чутлива до виявлення позитивних випадків (DELAY=1, прострочки немає).