LOGISTIC REGRESSION VERKHOVOD

# 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

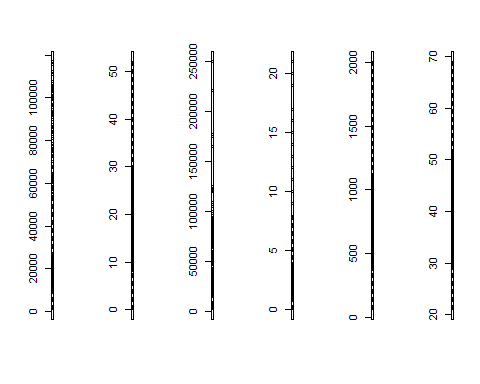
library(ggplot2)

## 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)

## Features Scaling

sc <- f[,c('LOAN\_AMOUNT','CLIENT\_TOTALEXPERIENCE', 'CLIENT\_TOGETHER.INCOME', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM', 'AGE')]   
sc <- scale(sc)   
f$LOAN\_AMOUNT <- sc[,c('LOAN\_AMOUNT')]   
f$CLIENT\_TOTALEXPERIENCE <- sc[,c('CLIENT\_TOTALEXPERIENCE')]   
f$CLIENT\_TOGETHER.INCOME <- sc[,c('CLIENT\_TOGETHER.INCOME')]   
f$LOAN\_OUTSTANDINGLOANSCOUNT <- sc[,c('LOAN\_OUTSTANDINGLOANSCOUNT')]   
f$LOAN.TERM <- sc[,c('LOAN.TERM')]   
f$AGE <- sc[,c('AGE')]   
head (f)

## LOAN\_AMOUNT BRANCH\_REGION CLIENT\_GENDER CLIENT\_FAMILYSTATUS  
## 1 -0.8167841 13 0 0  
## 2 -0.2172241 13 0 3  
## 3 -0.2883178 10 0 5  
## 4 -0.8518048 13 1 4  
## 5 -0.2172241 13 1 4  
## 6 -0.5591318 10 0 1  
## CLIENT\_TOTALEXPERIENCE CLIENT\_EDUCATION CLIENT\_ACTIVITYTYPE  
## 1 0.98171167 5 3  
## 2 -1.29541941 4 5  
## 3 0.08712446 0 5  
## 4 0.81905945 5 3  
## 5 0.90038556 5 5  
## 6 0.90038556 5 3  
## CLIENT\_TOGETHER.INCOME LOAN\_OUTSTANDINGLOANSCOUNT LOAN\_OVERDUE\_EXIST\_FLAG  
## 1 -0.6204727 -0.4216415 0  
## 2 -0.5420461 0.8928558 1  
## 3 -0.1443111 -1.2979730 0  
## 4 -0.7997336 -0.8598072 1  
## 5 -0.3683872 -0.4216415 0  
## 6 -0.6484823 -0.8598072 1  
## EMPLOYMENTTYPE LOAN.TERM ZODIAC ZODIAC\_CHINA AGE CHANGE.WORK  
## 1 1 -1.2599295 7 9 0.31818129 0  
## 2 1 0.8796904 1 10 -1.41516066 0  
## 3 4 -1.2599295 11 1 0.01673052 0  
## 4 0 0.8796904 11 9 0.31818129 0  
## 5 1 0.8796904 5 11 0.99644554 0  
## 6 1 -0.1915831 7 3 0.61963207 0  
## REAL\_ESTATE CAR DELAY  
## 1 1 1 1  
## 2 1 1 1  
## 3 0 1 1  
## 4 0 1 0  
## 5 0 1 1  
## 6 0 1 0

Висновок: було проведене попереднє шкалювання кількісних змінних.

# 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 (Benchmark model)

class\_lr <- glm(DELAY ~ ., f\_train, family = binomial)  
summary(class\_lr)

##   
## Call:  
## glm(formula = DELAY ~ ., family = binomial, data = f\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4550 -0.9644 0.4638 0.8958 2.2703   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.200442 0.282055 0.711 0.477302   
## LOAN\_AMOUNT -0.150129 0.064223 -2.338 0.019407 \*   
## BRANCH\_REGION 0.064421 0.008240 7.818 5.35e-15 \*\*\*  
## CLIENT\_GENDER -0.412213 0.085968 -4.795 1.63e-06 \*\*\*  
## CLIENT\_FAMILYSTATUS 0.182349 0.038075 4.789 1.67e-06 \*\*\*  
## CLIENT\_TOTALEXPERIENCE 0.365244 0.110523 3.305 0.000951 \*\*\*  
## CLIENT\_EDUCATION -0.104014 0.018650 -5.577 2.44e-08 \*\*\*  
## CLIENT\_ACTIVITYTYPE -0.091224 0.026282 -3.471 0.000519 \*\*\*  
## CLIENT\_TOGETHER.INCOME 0.008986 0.064343 0.140 0.888936   
## LOAN\_OUTSTANDINGLOANSCOUNT -0.411516 0.047830 -8.604 < 2e-16 \*\*\*  
## LOAN\_OVERDUE\_EXIST\_FLAG -0.141648 0.095045 -1.490 0.136138   
## EMPLOYMENTTYPE 0.136471 0.060885 2.241 0.024997 \*   
## LOAN.TERM -0.516033 0.049166 -10.496 < 2e-16 \*\*\*  
## ZODIAC -0.011627 0.011825 -0.983 0.325453   
## ZODIAC\_CHINA -0.004746 0.012156 -0.390 0.696242   
## AGE 0.325639 0.110565 2.945 0.003227 \*\*   
## CHANGE.WORK 0.328041 0.115440 2.842 0.004488 \*\*   
## REAL\_ESTATE 0.145820 0.089663 1.626 0.103882   
## CAR -0.427739 0.156248 -2.738 0.006189 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4362.0 on 3199 degrees of freedom  
## Residual deviance: 3558.8 on 3181 degrees of freedom  
## AIC: 3596.8  
##   
## Number of Fisher Scoring iterations: 4

Висновок: значущими змінними є BRANCH\_REGION, CLIENT\_GENDER, CLIENT\_FAMILYSTATUS, CLIENT\_TOTALEXPERIENCE, CLIENT\_EDUCATION,CLIENT\_ACTIVITYTYPE, LOAN\_OUTSTANDINGLOANSCOUNT та LOAN.TERM.

## Optimized model

class\_opt <- glm(DELAY ~ BRANCH\_REGION+CLIENT\_GENDER+CLIENT\_FAMILYSTATUS+CLIENT\_TOTALEXPERIENCE+CLIENT\_EDUCATION+CLIENT\_ACTIVITYTYPE+LOAN\_OUTSTANDINGLOANSCOUNT+LOAN.TERM, f\_train, family = binomial)  
summary(class\_opt)

##   
## Call:  
## glm(formula = DELAY ~ BRANCH\_REGION + CLIENT\_GENDER + CLIENT\_FAMILYSTATUS +   
## CLIENT\_TOTALEXPERIENCE + CLIENT\_EDUCATION + CLIENT\_ACTIVITYTYPE +   
## LOAN\_OUTSTANDINGLOANSCOUNT + LOAN.TERM, family = binomial,   
## data = f\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4762 -0.9787 0.4798 0.9031 2.2430   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.116914 0.193097 -0.605 0.544870   
## BRANCH\_REGION 0.067386 0.008141 8.278 < 2e-16 \*\*\*  
## CLIENT\_GENDER -0.430401 0.082238 -5.234 1.66e-07 \*\*\*  
## CLIENT\_FAMILYSTATUS 0.178147 0.037505 4.750 2.04e-06 \*\*\*  
## CLIENT\_TOTALEXPERIENCE 0.630014 0.044165 14.265 < 2e-16 \*\*\*  
## CLIENT\_EDUCATION -0.099724 0.017841 -5.590 2.28e-08 \*\*\*  
## CLIENT\_ACTIVITYTYPE -0.092375 0.025217 -3.663 0.000249 \*\*\*  
## LOAN\_OUTSTANDINGLOANSCOUNT -0.413691 0.042767 -9.673 < 2e-16 \*\*\*  
## LOAN.TERM -0.572440 0.043471 -13.168 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4362.0 on 3199 degrees of freedom  
## Residual deviance: 3597.1 on 3191 degrees of freedom  
## AIC: 3615.1  
##   
## Number of Fisher Scoring iterations: 4

Висновок: всі змінні оптимізованої моделі є значущими.

# Predicting

p <- predict(class\_opt, f\_test[, c('BRANCH\_REGION', 'CLIENT\_GENDER', 'CLIENT\_FAMILYSTATUS', 'CLIENT\_TOTALEXPERIENCE', 'CLIENT\_EDUCATION', 'CLIENT\_ACTIVITYTYPE', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM')], type = 'response')  
y <- ifelse(p > 0.5, 1, 0)

Висновок: розраховані ймовірності віднесення об’єктів до кожного з двох класів (вектор р), визначені класи об’єктів (вектор у).

## Confusion Matrix

cm = table(f\_test[, 'DELAY'], y > 0.5)  
print(cm)

##   
## FALSE TRUE  
## 0 203 136  
## 1 91 370

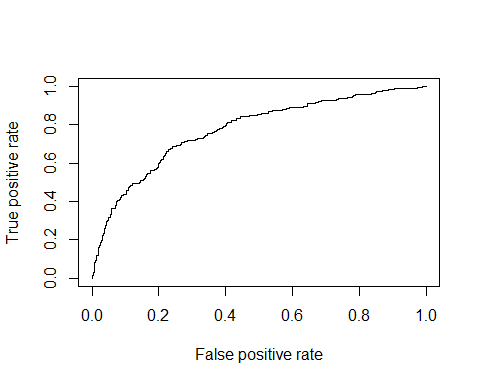
Висновок: точність моделі - 71,63 %, частка невірно класифікованих випадків – 28.38 %. Чутливість моделі – 80.26 %, специфічність – 59.88%, тобто модель більш чутлива до виявлення позитивних випадків (клієнтів, що не мають прострочки).

## ROC

library(ROCR)

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

pref <- prediction(p, f\_test$DELAY)  
perf <- performance(pref, "tpr", "fpr")  
plot(perf)

 Висновок: ROC-крива показує співвідношення істинно-позитивних і хибно-позитивних випадків. Вона свідчить про середню якість моделі.

# Write prepared data to the file

write.csv2(f\_train, file = "clients\_train.csv")  
write.csv2(f\_test, file = "clients\_test.csv")

Висновок: навчальна та тестова вибірки збережені в окремих файлах.

# K-Nearest Neighbors (K-NN)

# Fitting & predicting

library(class)  
y = knn(train = f\_train[,c( 'BRANCH\_REGION', 'CLIENT\_GENDER', 'CLIENT\_FAMILYSTATUS', 'CLIENT\_TOTALEXPERIENCE', 'CLIENT\_EDUCATION', 'CLIENT\_ACTIVITYTYPE', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM')],  
 test = f\_test[,c( 'BRANCH\_REGION', 'CLIENT\_GENDER', 'CLIENT\_FAMILYSTATUS', 'CLIENT\_TOTALEXPERIENCE', 'CLIENT\_EDUCATION', 'CLIENT\_ACTIVITYTYPE', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM')],  
 cl = f\_train[, 'DELAY'],  
 k = 30,  
 prob = TRUE)

Висновок: і навчання, і прогнозування за моделлю k найближчих сусідів здійснюється однією функцією. У результаті отримуємо вектор класів об’єктів.

## Confusion Matrix

cm1 = table(f\_test[, 'DELAY'], y == '1')  
print(cm1)

##   
## FALSE TRUE  
## 0 205 134  
## 1 83 378

Висновок: точність моделі – 72,5 %, частка невірно класифікованих випадків – 27,5 %. Чутливість – 82,43 %, специфічність – 59,00 %, тобто модель більш чутлива до виявлення позитивних випадків (кредиторів, що не мають прострочки).

# Fitting SVM model

# install.packages('e1071')  
library(e1071)

## Warning: package 'e1071' was built under R version 3.6.3

class\_svm\_l = svm(DELAY ~BRANCH\_REGION+CLIENT\_GENDER+CLIENT\_FAMILYSTATUS+CLIENT\_TOTALEXPERIENCE+CLIENT\_EDUCATION+CLIENT\_ACTIVITYTYPE+LOAN\_OUTSTANDINGLOANSCOUNT+LOAN.TERM, data = f\_train, kernel = 'linear')  
summary(class\_svm\_l)

##   
## Call:  
## svm(formula = DELAY ~ BRANCH\_REGION + CLIENT\_GENDER + CLIENT\_FAMILYSTATUS +   
## CLIENT\_TOTALEXPERIENCE + CLIENT\_EDUCATION + CLIENT\_ACTIVITYTYPE +   
## LOAN\_OUTSTANDINGLOANSCOUNT + LOAN.TERM, data = f\_train, kernel = "linear")  
##   
##   
## Parameters:  
## SVM-Type: eps-regression   
## SVM-Kernel: linear   
## cost: 1   
## gamma: 0.125   
## epsilon: 0.1   
##   
##   
## Number of Support Vectors: 2980

Висновок: для навчання базової моделі, заснованої на методі опорних векторів, вибрано лінійне ядро.

# Predicting

p <- predict(class\_svm\_l, f\_test[, c( 'BRANCH\_REGION', 'CLIENT\_GENDER', 'CLIENT\_FAMILYSTATUS', 'CLIENT\_TOTALEXPERIENCE', 'CLIENT\_EDUCATION', 'CLIENT\_ACTIVITYTYPE', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM')])  
y <- ifelse(p > 0.5, 1, 0)

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

## Confusion Matrix

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

## y  
## 0 1  
## 0 206 133  
## 1 97 364

Висновок: точність моделі – 71.25 %, частка невірно класифікованих випадків – 28.75 %. Чутливість 78.96 %, специфічність – 60.77 %, тобто модель більш чутлива до виявлення позитивних випадків (кредиторів, що не мають прострочки).

# Fitting RBF-kernel model

# install.packages('e1071')  
library(e1071)  
class\_svm\_r = svm(DELAY ~ BRANCH\_REGION+CLIENT\_GENDER+CLIENT\_FAMILYSTATUS+CLIENT\_TOTALEXPERIENCE+CLIENT\_EDUCATION+CLIENT\_ACTIVITYTYPE+LOAN\_OUTSTANDINGLOANSCOUNT+LOAN.TERM, data = f\_train, kernel = 'radial')  
summary(class\_svm\_r)

##   
## Call:  
## svm(formula = DELAY ~ BRANCH\_REGION + CLIENT\_GENDER + CLIENT\_FAMILYSTATUS +   
## CLIENT\_TOTALEXPERIENCE + CLIENT\_EDUCATION + CLIENT\_ACTIVITYTYPE +   
## LOAN\_OUTSTANDINGLOANSCOUNT + LOAN.TERM, data = f\_train, kernel = "radial")  
##   
##   
## Parameters:  
## SVM-Type: eps-regression   
## SVM-Kernel: radial   
## cost: 1   
## gamma: 0.125   
## epsilon: 0.1   
##   
##   
## Number of Support Vectors: 2439

Висновок: для навчання моделі, заснованої на методі опорних векторів, вибрано нелінійне ядро.

# Predicting

p <- predict(class\_svm\_r, f\_test[, c('BRANCH\_REGION', 'CLIENT\_GENDER', 'CLIENT\_FAMILYSTATUS', 'CLIENT\_TOTALEXPERIENCE', 'CLIENT\_EDUCATION', 'CLIENT\_ACTIVITYTYPE', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM')])  
y <- ifelse(p > 0.5, 1, 0)

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

## Confusion Matrix

cm2 = table(f\_test[, 'DELAY'], y)  
print(cm2)

## y  
## 0 1  
## 0 203 136  
## 1 87 374

Висновок: точність моделі 72.13 %, частка невірно класифікованих випадків – 27.88 %. Чутливість – 81.13 %, специфічність – 59.88 %, тобто модель більш чутлива до виявлення позитивних випадків (кредиторів, що не мають прострочки).

# Naive Bayes

# Fitting

# install.packages('e1071')  
library(e1071)  
f\_train$DELAY <- as.factor(f\_train$DELAY)  
f\_test$DELAY <- as.factor(f\_test$DELAY)  
class\_nb = naiveBayes(DELAY ~ BRANCH\_REGION + CLIENT\_GENDER + CLIENT\_FAMILYSTATUS +   
 CLIENT\_TOTALEXPERIENCE + CLIENT\_EDUCATION + CLIENT\_ACTIVITYTYPE + LOAN\_OUTSTANDINGLOANSCOUNT +   
 LOAN.TERM, data = f\_train)

Висновок: для навчання моделі використано функцію naiveBayes

# Predicting

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

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

## Confusion Matrix

cm3 = table(f\_test[, 'DELAY'], y)  
print(cm3)

## y  
## 0 1  
## 0 193 146  
## 1 94 367

Висновок: точність моделі – 70.00 %, частка невірно класифікованих випадків – 30.00 %. Чутливість – 79.61 %, специфічність – 56.93 %, тобто модель більш чутлива до виявлення позитивних випадків (кредиторів, що не мають прострочки).