LOGISTIC REGRESSION VERKHOVOD

# Download the data

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

## LOAN\_PRODUCT\_TYPE LOAN\_AMOUNT BRANCH\_REGION  
## 1 Cпоживчий кредит без довiдки про доходи 6224.07 Харків  
## 2 Cпоживчий кредит без довiдки про доходи 16816.14 Харків  
## 3 Cпоживчий кредит без довiдки про доходи 15560.17 Полтавська область  
## 4 Cпоживчий кредит без довiдки про доходи 5605.38 Харків  
## 5 Cпоживчий кредит без довiдки про доходи 16816.14 Харків  
## 6 Cпоживчий кредит без довiдки про доходи 10775.86 Полтавська область  
## CLIENT\_GENDER CLIENT\_FAMILYSTATUS CLIENT\_TOTALEXPERIENCE CLIENT\_EDUCATION  
## 1 Жінка В громадянскому шлюбі 31 Середня - технічна  
## 2 Жінка Неодружений(-на) 3 Середня  
## 3 Жінка Розлучений (-на) 20 Вища освіта  
## 4 Чоловік Одружений (-на) 29 Середня - технічна  
## 5 Чоловік Одружений (-на) 30 Середня - технічна  
## 6 Жінка Вдова (-ець) 30 Середня - технічна  
## CLIENT\_ACTIVITYTYPE CLIENT\_TOGETHER.INCOME LOAN\_OUTSTANDINGLOANSCOUNT  
## 1 інша 3750 2  
## 2 торгівля та послуги 4450 5  
## 3 торгівля та послуги 8000 0  
## 4 інша 2150 1  
## 5 торгівля та послуги 6000 2  
## 6 інша 3500 1  
## LOAN\_OVERDUE\_EXIST\_FLAG EMPLOYMENTTYPE LOAN.TERM ZODIAC  
## 1 Ні Офіційно працевлаштований 364 рак  
## 2 Так Офіційно працевлаштований 1095 весы  
## 3 Ні ФОП 364 телец  
## 4 Так Не працює 1095 телец  
## 5 Ні Офіційно працевлаштований 1095 лев  
## 6 Так Офіційно працевлаштований 729 рак  
## ZODIAC\_CHINA AGE CHANGE.WORK REAL\_ESTATE  
## 1 Свинья 45 була зміна місця роботи не має власної нерухомості  
## 2 Собака 22 була зміна місця роботи не має власної нерухомості  
## 3 Дракон 41 була зміна місця роботи має власну нерухомість  
## 4 Свинья 45 була зміна місця роботи має власну нерухомість  
## 5 Тигр 54 була зміна місця роботи має власну нерухомість  
## 6 Коза 49 була зміна місця роботи має власну нерухомість  
## CAR DELAY  
## 1 не має власного нет прострочки  
## 2 не має власного нет прострочки  
## 3 не має власного нет прострочки  
## 4 не має власного есть прострочка  
## 5 не має власного нет прострочки  
## 6 не має власного есть прострочка

Висновок: для побудови моделі банківського скорингу 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\_PRODUCT\_TYPE\* 1 4000 1.74 1.56 1.00 1.32  
## LOAN\_AMOUNT 2 4000 21470.59 20885.55 16816.14 17493.79  
## BRANCH\_REGION\* 3 4000 11.22 4.97 12.00 11.48  
## CLIENT\_GENDER\* 4 4000 1.49 0.50 1.00 1.49  
## CLIENT\_FAMILYSTATUS\* 5 4000 4.54 1.11 5.00 4.68  
## CLIENT\_TOTALEXPERIENCE 6 4000 18.93 12.30 17.00 18.24  
## CLIENT\_EDUCATION\* 7 4000 3.75 2.31 5.00 3.81  
## CLIENT\_ACTIVITYTYPE\* 8 4000 4.32 1.59 4.00 4.42  
## CLIENT\_TOGETHER.INCOME 9 4000 10062.58 14185.66 7000.00 7404.52  
## LOAN\_OUTSTANDINGLOANSCOUNT 10 4000 3.01 2.48 2.00 2.67  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 11 4000 1.69 0.46 2.00 1.73  
## EMPLOYMENTTYPE\* 12 4000 2.22 0.84 2.00 2.00  
## LOAN.TERM 13 4000 794.45 341.65 729.00 803.05  
## ZODIAC\* 14 4000 6.25 3.44 6.00 6.21  
## ZODIAC\_CHINA\* 15 4000 6.62 3.43 7.00 6.65  
## AGE 16 4000 40.78 13.27 40.00 40.25  
## CHANGE.WORK\* 17 4000 1.16 0.36 1.00 1.07  
## REAL\_ESTATE\* 18 4000 1.39 0.49 1.00 1.36  
## CAR\* 19 4000 1.91 0.29 2.00 2.00  
## DELAY\* 20 4000 1.58 0.49 2.00 1.60  
## mad min max range skew kurtosis  
## LOAN\_PRODUCT\_TYPE\* 0.00 1.0 6.0 5.0 1.93 2.19  
## 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\_PRODUCT\_TYPE\* 0.02  
## 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 – 20, з них якісних – 14, кількісних – 6. Пропущених значень немає. Викиди мають змінні: LOAN\_AMOUNT, CLIENT\_TOGETHER.INCOME, LOAN\_OUTSTANDINGLOANSCOUNT

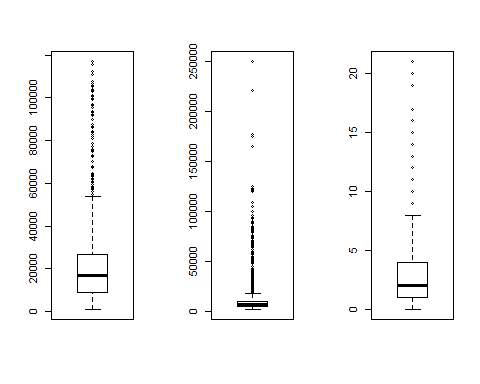
library(ggplot2)

## 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(1,3))  
boxplot(f$LOAN\_AMOUNT)  
boxplot(f$CLIENT\_TOGETHER.INCOME)  
boxplot(f$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\_PRODUCT\_TYPE\* 1 4000 1.74 1.56 1.00 1.32  
## LOAN\_AMOUNT 2 4000 20653.71 17666.41 16816.14 17493.79  
## BRANCH\_REGION\* 3 4000 11.22 4.97 12.00 11.48  
## CLIENT\_GENDER\* 4 4000 1.49 0.50 1.00 1.49  
## CLIENT\_FAMILYSTATUS\* 5 4000 4.54 1.11 5.00 4.68  
## CLIENT\_TOTALEXPERIENCE 6 4000 18.93 12.30 17.00 18.24  
## CLIENT\_EDUCATION\* 7 4000 3.75 2.31 5.00 3.81  
## CLIENT\_ACTIVITYTYPE\* 8 4000 4.32 1.59 4.00 4.42  
## CLIENT\_TOGETHER.INCOME 9 4000 9288.06 8925.54 7000.00 7404.52  
## LOAN\_OUTSTANDINGLOANSCOUNT 10 4000 2.96 2.28 2.00 2.67  
## LOAN\_OVERDUE\_EXIST\_FLAG\* 11 4000 1.69 0.46 2.00 1.73  
## EMPLOYMENTTYPE\* 12 4000 2.22 0.84 2.00 2.00  
## LOAN.TERM 13 4000 794.45 341.65 729.00 803.05  
## ZODIAC\* 14 4000 6.25 3.44 6.00 6.21  
## ZODIAC\_CHINA\* 15 4000 6.62 3.43 7.00 6.65  
## AGE 16 4000 40.78 13.27 40.00 40.25  
## CHANGE.WORK\* 17 4000 1.16 0.36 1.00 1.07  
## REAL\_ESTATE\* 18 4000 1.39 0.49 1.00 1.36  
## CAR\* 19 4000 1.91 0.29 2.00 2.00  
## DELAY\* 20 4000 1.58 0.49 2.00 1.60  
## mad min max range skew kurtosis  
## LOAN\_PRODUCT\_TYPE\* 0.00 1.0 6.00 5.00 1.93 2.19  
## 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\_PRODUCT\_TYPE\* 0.02  
## 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$LOAN\_PRODUCT\_TYPE <- as.numeric(as.factor(f$LOAN\_PRODUCT\_TYPE))-1  
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.

**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\_PRODUCT\_TYPE LOAN\_AMOUNT BRANCH\_REGION CLIENT\_GENDER CLIENT\_FAMILYSTATUS  
## 1 0 -0.8167841 13 0 0  
## 2 0 -0.2172241 13 0 3  
## 3 0 -0.2883178 10 0 5  
## 4 0 -0.8518048 13 1 4  
## 5 0 -0.2172241 13 1 4  
## 6 0 -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   
## -2.7504 -0.9288 0.4372 0.8840 2.3853   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.136788 0.284642 0.481 0.63083   
## LOAN\_PRODUCT\_TYPE 0.284191 0.038652 7.352 1.95e-13 \*\*\*  
## LOAN\_AMOUNT -0.418165 0.074282 -5.629 1.81e-08 \*\*\*  
## BRANCH\_REGION 0.061477 0.008332 7.378 1.60e-13 \*\*\*  
## CLIENT\_GENDER -0.438797 0.086916 -5.049 4.45e-07 \*\*\*  
## CLIENT\_FAMILYSTATUS 0.181909 0.038576 4.716 2.41e-06 \*\*\*  
## CLIENT\_TOTALEXPERIENCE 0.362602 0.111466 3.253 0.00114 \*\*   
## CLIENT\_EDUCATION -0.098081 0.018814 -5.213 1.86e-07 \*\*\*  
## CLIENT\_ACTIVITYTYPE -0.096434 0.026608 -3.624 0.00029 \*\*\*  
## CLIENT\_TOGETHER.INCOME -0.003615 0.064687 -0.056 0.95543   
## LOAN\_OUTSTANDINGLOANSCOUNT -0.451045 0.048760 -9.250 < 2e-16 \*\*\*  
## LOAN\_OVERDUE\_EXIST\_FLAG -0.144587 0.096045 -1.505 0.13222   
## EMPLOYMENTTYPE 0.089457 0.060739 1.473 0.14080   
## LOAN.TERM -0.488819 0.049653 -9.845 < 2e-16 \*\*\*  
## ZODIAC -0.011048 0.011940 -0.925 0.35481   
## ZODIAC\_CHINA -0.005207 0.012282 -0.424 0.67157   
## AGE 0.277289 0.111549 2.486 0.01293 \*   
## CHANGE.WORK 0.267382 0.116907 2.287 0.02219 \*   
## REAL\_ESTATE 0.142171 0.090584 1.569 0.11654   
## CAR -0.437201 0.157788 -2.771 0.00559 \*\*   
## ---  
## 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: 3501.3 on 3180 degrees of freedom  
## AIC: 3541.3  
##   
## Number of Fisher Scoring iterations: 4

Висновок: значущими змінними є LOAN\_PRODUCT\_TYPE, LOAN\_AMOUNT, BRANCH\_REGION, CLIENT\_GENDER, CLIENT\_FAMILYSTATUS, CLIENT\_EDUCATION, LOAN\_OUTSTANDINGLOANSCOUNT та LOAN.TERM.

# Optimized model

class\_opt <- glm(DELAY ~ LOAN\_PRODUCT\_TYPE+LOAN.TERM, f\_train, family = binomial)  
summary(class\_opt)

##   
## Call:  
## glm(formula = DELAY ~ LOAN\_PRODUCT\_TYPE + LOAN.TERM, family = binomial,   
## data = f\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1317 -1.0427 0.7536 1.0239 1.5043   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.24726 0.04108 6.019 1.76e-09 \*\*\*  
## LOAN\_PRODUCT\_TYPE 0.14174 0.02536 5.589 2.28e-08 \*\*\*  
## LOAN.TERM -0.65332 0.04032 -16.203 < 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: 4067.5 on 3197 degrees of freedom  
## AIC: 4073.5  
##   
## Number of Fisher Scoring iterations: 4

Висновок: всі змінні оптимізованої моделі є значущими.

# Predicting

p <- predict(class\_opt, f\_test[, c('LOAN\_PRODUCT\_TYPE', '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)

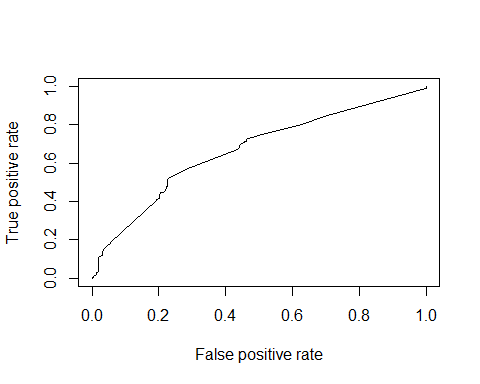
##   
## FALSE TRUE  
## 0 183 156  
## 1 133 328

Висновок: точність моделі - (183 + 156) / 800 = 63,88 %, частка невірно класифікованих випадків – (133 + 156) / 800 = 36,13 %. Чутливість моделі – 328 / (133 + 328) = 71,15 %, специфічність – 183 / (183 + 156) = 53,98%, тобто модель більш чутлива до виявлення позитивних випадків (клієнтів, що мають заборгованість). ## ROC

library(ROCR)

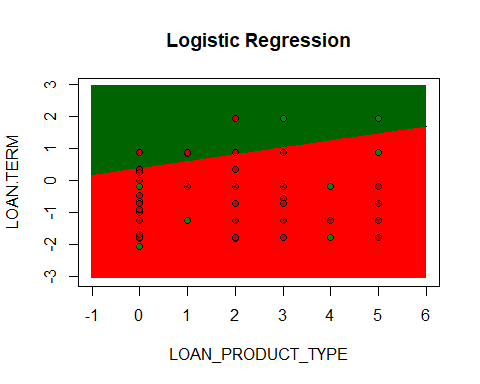
## Warning: package 'ROCR' was built under R version 3.6.3

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

 Висновок: співвідношення істинно-позитивних і хибно-позитивних випадків свідчить про середню якість моделі.

# Visualising the Test set results

library(ggplot2)  
set = f\_test[,c('LOAN\_PRODUCT\_TYPE', 'LOAN.TERM','DELAY')]  
X1 = seq(min(set['LOAN\_PRODUCT\_TYPE']) - 1, max(set['LOAN\_PRODUCT\_TYPE']) + 1, by = 0.01)  
X2 = seq(min(set['LOAN.TERM']) - 1, max(set['LOAN.TERM']) + 1, by = 0.01)  
grid\_set = expand.grid(X1, X2)  
colnames(grid\_set) = c('LOAN\_PRODUCT\_TYPE', 'LOAN.TERM')  
prob\_set = predict(class\_opt, grid\_set, type = 'response')  
y\_grid = ifelse(prob\_set > 0.5, 1, 0)  
plot(set[, -3],  
 main = 'Logistic Regression',  
 xlab = 'LOAN\_PRODUCT\_TYPE', ylab = 'LOAN.TERM',  
 xlim = range(X1), ylim = range(X2))  
contour(X1, X2, matrix(as.numeric(y\_grid), length(X1), length(X2)), add = TRUE)  
points(grid\_set, pch = '.', col = ifelse(y\_grid == 1, 'red', 'darkgreen'))  
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'red3', 'green4'))



Висновок: на графіку червоним позначені випадки затримки повернення кредиту, зеленим – хороші кредитори. Червоним виділена зона високої ймовірності неповернення кредиту. Модель описує лінійний варіант розділяючої кривої.

# 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.