Naive Bayes

# Download the data

set.seed(123)  
setwd('C:/Users/VerkhovodTS/Desktop/R')  
f\_train <- read.csv2('clients\_train.csv', header = TRUE, encoding = 'UNICOD')  
f\_test <- read.csv2('clients\_test.csv', header = TRUE, encoding = 'UNICOD')  
f\_train <- f\_train[-1]  
f\_test <- f\_test[-1]

# Fitting

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

## Warning: package 'e1071' 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\_nb = naiveBayes(DELAY ~ LOAN\_PRODUCT\_TYPE + LOAN.TERM, data = f\_train)

Висновок: для навчання моделі використано функцію naiveBayes.

# Predicting

y <- predict(class\_nb, f\_test[, c('LOAN\_PRODUCT\_TYPE','LOAN.TERM')])

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

# Confusion Matrix

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

## y  
## 0 1  
## 0 196 143  
## 1 153 308

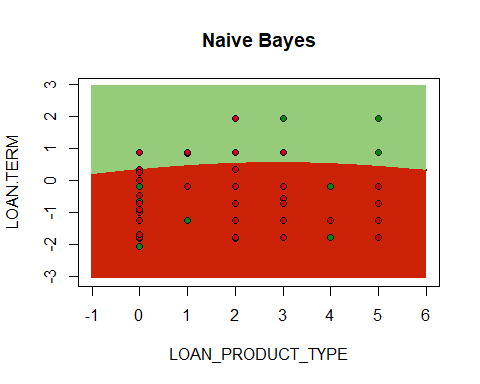
Висновок: точність моделі – (196+308) / 800 = 63,00 %, частка невірно класифікованих випадків – (153+143) / 800 = 37,00 %. Чутливість – 308 / (153+308) = 66,81 %, специфічність – 196 / (196+143) = 57,82 %, тобто модель більш чутлива до виявлення позитивних випадків (кредиторів, що мають заборгованість).

**Visualising the Test set results**

library(ggplot2)

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

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')  
y\_grid = predict(class\_nb, grid\_set)  
plot(set[, -3],  
 main = 'Naive Bayes',  
 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, '#cc2208', '#94cc7a'))  
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'red3', 'green4'))



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