DZ3

Stepan

13 10 2020

# Download the data

#Download the files  
setwd('C:/Users/Stepan/Desktop/6 курс/Машинное обучение/DZ3')  
d\_train <- read.csv2('DATASET\_train.csv', header = TRUE, encoding = 'UNICOD')  
d\_train <- d\_train[,-1]  
  
d\_test <- read.csv2('DATASET\_test.csv', header = TRUE, encoding = 'UNICOD')  
d\_test <- d\_test[,-1]

# Decision Tree Regression

## Fitting simple tree

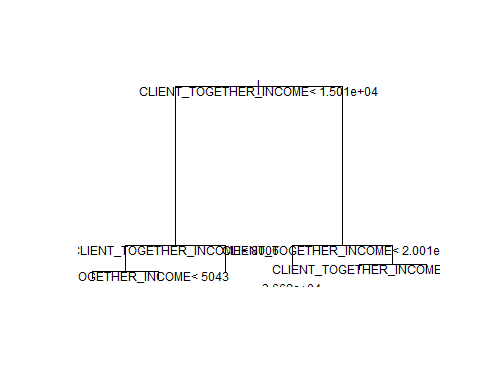
#install.packages('rpart')  
library(rpart)

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

dt <- rpart(LOAN\_AMOUNT ~ CLIENT\_TOGETHER\_INCOME, d\_train, control = rpart.control(minsplit = 50))  
plot(dt)  
text(dt, pos = 1, cex = .75, col = 1, font = 1)

## #Висновок: окремо завантажені навчальна і тестова вибірки.

## #Висновок: окремо завантажені навчальна і тестова вибірки.

 ## Predicting

p\_dt <- predict(dt, d\_test)  
  
train\_mse\_dt <- sum((d\_train$LOAN\_AMOUNT-predict(dt, d\_train))^2) /length(d\_train$LOAN\_AMOUNT)  
test\_mse\_dt <- sum((d\_test$LOAN\_AMOUNT-p\_dt)^2)/length(p\_dt)  
  
train\_mse\_dt

## [1] 190234536

test\_mse\_dt

## [1] 189282111

**#Висновок: значення середньоквадратичної помилки навчальній вибірці – 190234536, на тестовій вибірці – 189282111, таким чинном перенавчання немає.**

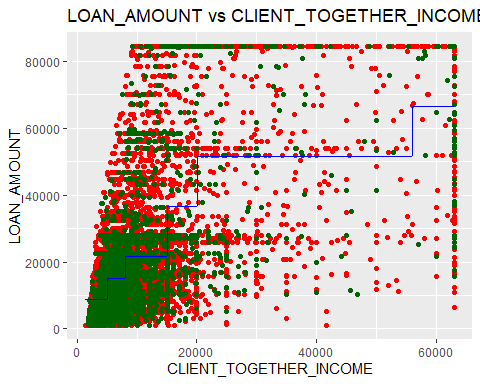
## Visualising

library(ggplot2)

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

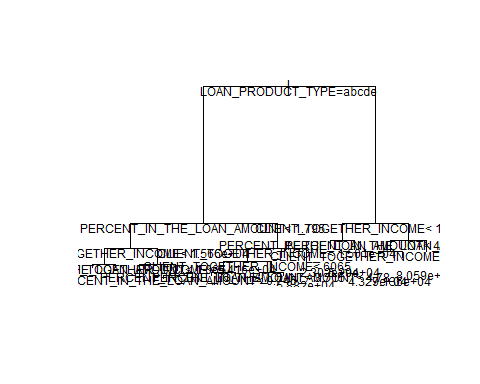
x\_grid <- seq(min(d\_train$CLIENT\_TOGETHER\_INCOME), max(d\_train$CLIENT\_TOGETHER\_INCOME), 0.01)  
ggplot() +  
 geom\_point(aes(d\_train$CLIENT\_TOGETHER\_INCOME, d\_train$LOAN\_AMOUNT),colour = 'red') +  
 geom\_point(aes(d\_test$CLIENT\_TOGETHER\_INCOME, d\_test$LOAN\_AMOUNT),colour = 'dark green') +  
 geom\_line(aes(x\_grid, predict(dt, data.frame(CLIENT\_TOGETHER\_INCOME = x\_grid))),colour = 'blue') +  
 ggtitle('LOAN\_AMOUNT vs CLIENT\_TOGETHER\_INCOME') +  
 xlab('CLIENT\_TOGETHER\_INCOME') +  
 ylab('LOAN\_AMOUNT')

**#Висновок: на графіку червоним позначені точки навчальної вибірки, зеленим – точки тестової вибірки, синім – модельні значення.**



**## Fitting full tree**

# install.packages('rpart')  
library(rpart)  
ft <- rpart(LOAN\_AMOUNT ~ LOAN\_PRODUCT\_TYPE + BRANCH\_REGION + CLIENT\_GENDER + CLIENT\_TOTALEXPERIENCE + CLIENT\_EDUCATION + CLIENT\_TOGETHER\_INCOME + LOAN\_EXISTINGCUSTOMERFLAG + LOAN\_OVERDUE\_EXIST\_FLAG + SCR\_WORKPLACECONFIRMED + AGE + MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE + FAMILY\_PHONE + CAR + PERCENT\_IN\_THE\_LOAN\_AMOUNT + DELAY, d\_train, control = rpart.control(minsplit = 2))  
plot(ft)  
text(ft, pos = 1, cex = .75, col = 1, font = 1)



## Predicting

p\_ft <- predict(ft, d\_test)  
  
train\_mse\_ft <- sum((d\_train$LOAN\_AMOUNT-predict(dt, d\_train))^2)/length(d\_train$LOAN\_AMOUNT)  
test\_mse\_ft <- sum((d\_test$LOAN\_AMOUNT-p\_ft)^2)/length(p\_ft)  
  
train\_mse\_ft

## [1] 190234536

test\_mse\_ft

## [1] 51351540

**#Висновок: значення середньоквадратичної помилки трохи зросли на навчальній вибірці – 190234536, на тестовій вибірці – 51351540, тобто перенавчання немає.**

# Random forest

## Fitting

# 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

set.seed(1234)  
rf = randomForest(x = d\_train['CLIENT\_TOGETHER\_INCOME'],  
 y = d\_train$LOAN\_AMOUNT,  
 ntree = 5)

## Predicting

p\_rf <- predict(rf, d\_test)  
  
train\_mse\_rf <- sum((d\_train$LOAN\_AMOUNT-predict(rf, d\_train))^2)/length(d\_train$LOAN\_AMOUNT)  
test\_mse\_rf <- sum((d\_test$LOAN\_AMOUNT-p\_rf)^2)/length(p\_rf)  
  
train\_mse\_rf

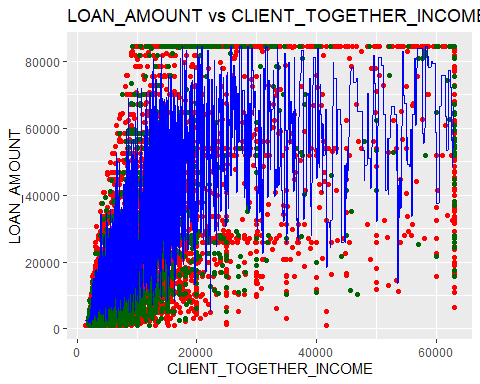
## [1] 110970118

test\_mse\_rf

## [1] 198435544

## Visualising

ggplot() +  
 geom\_point(aes(d\_train$CLIENT\_TOGETHER\_INCOME, d\_train$LOAN\_AMOUNT),colour = 'red') +  
 geom\_point(aes(d\_test$CLIENT\_TOGETHER\_INCOME, d\_test$LOAN\_AMOUNT),colour = 'dark green') +  
 geom\_line(aes(x\_grid, predict(rf, data.frame(CLIENT\_TOGETHER\_INCOME = x\_grid))),colour = 'blue') +  
 ggtitle('LOAN\_AMOUNT vs CLIENT\_TOGETHER\_INCOME') +  
 xlab('CLIENT\_TOGETHER\_INCOME') +  
 ylab('LOAN\_AMOUNT')



**#Висновок: на графіку червоним позначені точки навчальної вибірки, зеленим – точки тестової вибірки, синім – модельні значення.**

# Saving results

fit <- read.csv2('LOAN\_AMOUNT.csv', header = TRUE, encoding = 'UNICOD')  
fit$p\_ft <- p\_ft  
fit$p\_rf <- p\_rf  
head(fit)

## X p\_sr p\_mr p\_pr p\_ft p\_rf  
## 1 1 14897.72 16354.910 14419.56 11719.648 12518.611  
## 2 2 12666.07 13736.473 13013.55 11719.648 8410.632  
## 3 3 14005.06 11064.337 13797.27 11719.648 10865.871  
## 4 4 19361.02 15610.066 18608.73 11254.198 17571.975  
## 5 5 14674.55 19378.418 14256.73 11719.648 11392.028  
## 6 6 19249.43 9929.325 18484.02 5430.537 19955.960

write.csv2(fit[-1], file = "DATASET\_fit.csv")

**#Висновок: результати моделювання збережені у файлі.**