DZ4

Stepan

13 10 2020

#Download the files  
setwd('C:/Users/Stepan/Desktop/6 курс/Машинное обучение/DZ4')  
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]

library (dplyr)

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

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

d\_train\_nn <- select(d\_train, LOAN\_AMOUNT,   
 CLIENT\_COUNTDEPENDENTS,   
 CLIENT\_TOTALEXPERIENCE,   
 CLIENT\_LASTEXPERIENCE,   
 CLIENT\_TOGETHER\_INCOME,   
 LOAN\_OUTSTANDINGLOANSCOUNT,   
 LOAN\_TERM,   
 AGE,   
 PERCENT\_IN\_THE\_LOAN\_AMOUNT)  
d\_test\_nn <- select(d\_test, LOAN\_AMOUNT,   
 CLIENT\_COUNTDEPENDENTS,   
 CLIENT\_TOTALEXPERIENCE,   
 CLIENT\_LASTEXPERIENCE,   
 CLIENT\_TOGETHER\_INCOME,   
 LOAN\_OUTSTANDINGLOANSCOUNT,   
 LOAN\_TERM,   
 AGE,   
 PERCENT\_IN\_THE\_LOAN\_AMOUNT)  
d\_train\_sc <- as.data.frame(scale(d\_train\_nn))  
d\_test\_sc <- as.data.frame(scale(d\_test\_nn))  
head (d\_train\_sc)

## LOAN\_AMOUNT CLIENT\_COUNTDEPENDENTS CLIENT\_TOTALEXPERIENCE  
## 1 -0.7537971 -0.5907210 0.83807529  
## 2 -0.1716522 -0.5907210 -1.43572907  
## 3 -0.2406809 -0.5907210 -0.05520499  
## 4 -0.7878006 -0.5907210 0.67566070  
## 5 -0.5036287 -0.5907210 0.75686799  
## 6 -0.1716522 0.9107161 -1.27331447  
## CLIENT\_LASTEXPERIENCE CLIENT\_TOGETHER\_INCOME LOAN\_OUTSTANDINGLOANSCOUNT  
## 1 0.5022747 -0.6728808 0.3902244  
## 2 -0.9077613 -0.6032891 2.5364566  
## 3 0.5022747 -0.2503599 -1.0405971  
## 4 0.5641184 -0.8319475 -0.3251863  
## 5 1.8628357 -0.6977350 -0.3251863  
## 6 -0.8582863 -0.4114149 0.3902244  
## LOAN\_TERM AGE PERCENT\_IN\_THE\_LOAN\_AMOUNT  
## 1 -0.98428097 0.1432336 -0.3182911  
## 2 1.11038632 -1.5901418 1.2190112  
## 3 -0.98428097 -0.1582230 -0.1079998  
## 4 1.11038632 0.1432336 0.3705943  
## 5 0.06161993 0.4446902 0.7114113  
## 6 1.11038632 -1.5901418 0.3923486

#install.packages('reshape')  
library(nnet)

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

dd\_ap <- nnet (LOAN\_AMOUNT ~  
 CLIENT\_COUNTDEPENDENTS+   
 CLIENT\_TOTALEXPERIENCE+   
 CLIENT\_LASTEXPERIENCE+   
 CLIENT\_TOGETHER\_INCOME+   
 LOAN\_OUTSTANDINGLOANSCOUNT+   
 LOAN\_TERM+   
 AGE+   
 PERCENT\_IN\_THE\_LOAN\_AMOUNT, d\_train\_sc, linout = TRUE ,size = 4, maxit = 10000)  
library(graphics)  
source(file = 'plot.nnet.R')  
plot.nnet(dd\_ap)

## Loading required package: scales

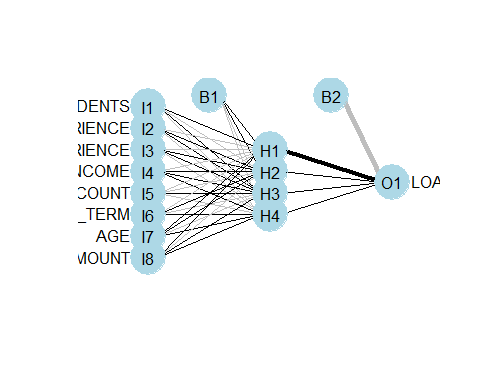
## Warning: package 'scales' was built under R version 3.6.3

## Loading required package: reshape

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

##   
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':  
##   
## rename



## Prediction

p\_y\_train <- predict(dd\_ap, d\_train\_sc)  
p\_y\_test <- predict(dd\_ap, d\_test\_sc)

## Invert the effect of the scale function

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

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

## Loading required package: lattice

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

sc\_y\_train <- scale(d\_train\_sc$LOAN\_AMOUNT)  
d\_train\_nn <- DMwR::unscale(p\_y\_train, sc\_y\_train)  
sc\_y\_test <- scale(d\_test\_sc$LOAN\_AMOUNT)  
d\_test\_nn <- DMwR::unscale(p\_y\_test, sc\_y\_test)

## MSE

train\_mse <- sum((d\_train\_sc$LOAN\_AMOUNT-d\_train\_sc)^2)/length(d\_train\_sc$LOAN\_AMOUNT)  
test\_mse <- sum((d\_test\_sc$LOAN\_AMOUNT-d\_test\_sc)^2)/length(d\_test\_sc$LOAN\_AMOUNT)  
train\_mse

## [1] 12.94963

test\_mse

## [1] 12.84013

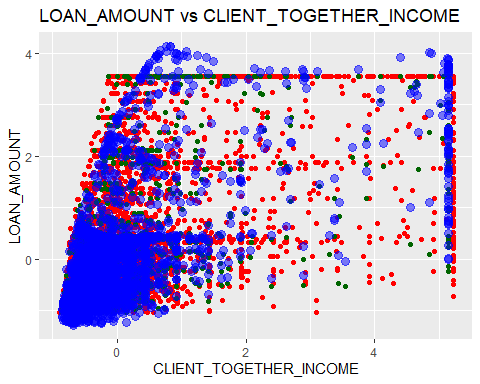
**#Висновок: значення середньоквадратичної помилки на навчальній вибірці – 12.94963, на тестовій вибірці – 12.84013, тобто перенавчання немає.**

library(ggplot2)

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

ggplot() +  
 geom\_point(aes(d\_train\_sc$CLIENT\_TOGETHER\_INCOME, d\_train\_sc$LOAN\_AMOUNT),colour = 'red') +  
 geom\_point(aes(d\_test\_sc$CLIENT\_TOGETHER\_INCOME, d\_test\_sc$LOAN\_AMOUNT),colour = 'dark green') +  
 geom\_point(aes(d\_test\_sc$CLIENT\_TOGETHER\_INCOME, d\_test\_nn),colour = 'blue', size = 3, alpha=0.5) +  
 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.**



# Fit NN-2

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

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

##   
## Attaching package: 'neuralnet'

## The following object is masked from 'package:dplyr':  
##   
## compute

# fit neural network  
nn = neuralnet(LOAN\_AMOUNT ~  
 CLIENT\_COUNTDEPENDENTS+   
 CLIENT\_TOTALEXPERIENCE+   
 CLIENT\_TOGETHER\_INCOME+   
 LOAN\_OUTSTANDINGLOANSCOUNT+   
 LOAN\_TERM, d\_train\_sc, hidden = 3, linear.output = T)  
# plot neural network  
plot(nn)

p\_y\_train\_nn <- predict(nn, d\_train\_sc)  
p\_y\_test\_nn <- predict(nn, d\_test\_sc)

## Invert the effect of the scale function

y\_train\_nn <- DMwR::unscale(p\_y\_train\_nn, sc\_y\_train)  
y\_test\_nn <- DMwR::unscale(p\_y\_test\_nn, sc\_y\_test)

## MSE

train\_mse\_nn <- sum((d\_train\_sc$LOAN\_AMOUNT-y\_train\_nn)^2)/length(d\_train\_sc$LOAN\_AMOUNT)  
test\_mse\_nn <- sum((d\_test\_sc$LOAN\_AMOUNT-y\_test\_nn)^2)/length(d\_test\_sc$LOAN\_AMOUNT)  
train\_mse\_nn

## [1] 0.4002886

test\_mse\_nn

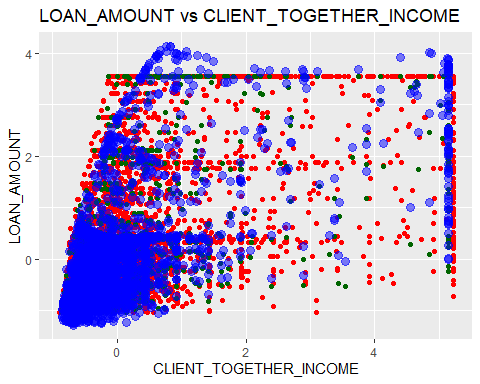
## [1] 0.399465

**#Висновок: значення середньоквадратичної помилки зменшились на навчальній вибірці – 0.4002886, на тестовій вибірці – 0.399465, тобто перенавчання немає.**

## Visualising

library(ggplot2)  
ggplot() +  
 geom\_point(aes(d\_train\_sc$CLIENT\_TOGETHER\_INCOME, d\_train\_sc$LOAN\_AMOUNT),colour = 'red') +  
 geom\_point(aes(d\_test\_sc$CLIENT\_TOGETHER\_INCOME, d\_test\_sc$LOAN\_AMOUNT),colour = 'dark green') +  
 geom\_point(aes(d\_test\_sc$CLIENT\_TOGETHER\_INCOME, d\_test\_nn),colour = 'blue', size = 3, alpha=0.5) +  
 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.**



fit <- read.csv2('DATASET\_fit.csv', header = TRUE, encoding = 'UNICOD')  
fit$nn <- y\_test\_nn  
head(fit)

## X p\_sr p\_mr p\_pr p\_ft p\_rf nn  
## 1 1 14897.72 16354.910 14419.56 11719.648 12518.611 -0.131584948  
## 2 2 12666.07 13736.473 13013.55 11719.648 8410.632 -0.290397688  
## 3 3 14005.06 11064.337 13797.27 11719.648 10865.871 -0.198134642  
## 4 4 19361.02 15610.066 18608.73 11254.198 17571.975 0.007642399  
## 5 5 14674.55 19378.418 14256.73 11719.648 11392.028 -0.337598037  
## 6 6 19249.43 9929.325 18484.02 5430.537 19955.960 -0.457279574

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

# #Висновок: результати моделювання збережені у файлі.