Linear Regression

Stepan

25 09 2020

#Source files are here  
setwd('C:/Users/Stepan/Desktop/6 курс/Машинное обучение/DZ2')  
##Features scaling is included in the packages we will work with  
#Download the files  
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]

#Висновок: окремо задані навчальна і тестова вибірки, видалені перші стовпчики з індексами об’єктів до кожної з підвибірок.

model\_sr <- lm(LOAN\_AMOUNT ~ CLIENT\_TOGETHER\_INCOME, d\_train)  
summary(model\_sr)

##   
## Call:  
## lm(formula = LOAN\_AMOUNT ~ CLIENT\_TOGETHER\_INCOME, data = d\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -71933 -8394 -2972 6015 66013   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.203e+03 1.768e+02 46.40 <2e-16 \*\*\*  
## CLIENT\_TOGETHER\_INCOME 1.116e+00 1.215e-02 91.85 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14320 on 13734 degrees of freedom  
## Multiple R-squared: 0.3805, Adjusted R-squared: 0.3805   
## F-statistic: 8436 on 1 and 13734 DF, p-value: < 2.2e-16

#Висновок: обрана змінна незначуща, коефіцієнт детермінації 0,38.

p\_sr <- predict(model\_sr, d\_test)  
r2\_sr <- 1-sum((d\_train$LOAN\_AMOUNT - predict(model\_sr,  
d\_train))^2)/sum((d\_train$LOAN\_AMOUNT - mean(d\_train$LOAN\_AMOUNT))^2)  
R2\_sr <- cor(d\_train$LOAN\_AMOUNT, fitted(model\_sr))^2 #simplier ex.  
train\_mse\_sr <- sum((d\_train$LOAN\_AMOUNT-predict(model\_sr,  
d\_train))^2)/length(d\_train$LOAN\_AMOUNT)  
test\_mse\_sr <- sum((d\_test$LOAN\_AMOUNT-p\_sr)^2)/length(p\_sr)  
r2\_sr

## [1] 0.3805169

R2\_sr

## [1] 0.3805169

train\_mse\_sr

## [1] 205067756

test\_mse\_sr

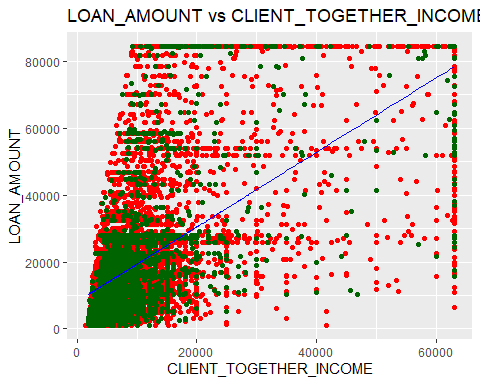
## [1] 199990348

#Висновок: вручну розраховані коефіцієнти детермінації. Значення середньоквадратичної похибки на навчальній вибірці – 205 067 756, на тестовій вибірці – 199 990 348, тобто перенавчання немає.

library(ggplot2)

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

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(d\_test$CLIENT\_TOGETHER\_INCOME, p\_sr),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..

model\_mr <- lm(data = d\_train, LOAN\_AMOUNT ~ .)  
summary(model\_mr)

##   
## Call:  
## lm(formula = LOAN\_AMOUNT ~ ., data = d\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60243 -2470 313 2883 33311   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 1.231e+07 3.754e+06  
## LOAN\_PRODUCT\_TYPE 1.918e+03 4.269e+01  
## BRANCH\_REGION -4.593e+01 1.318e+01  
## CLIENT\_GENDER 7.991e+02 1.286e+02  
## CLIENT\_FAMILYSTATUS -6.265e+01 5.053e+01  
## CLIENT\_COUNTDEPENDENTS 7.187e+01 9.565e+01  
## CLIENT\_TOTALEXPERIENCE 4.487e+01 1.265e+01  
## CLIENT\_LASTEXPERIENCE 1.634e+01 9.764e+00  
## CLIENT\_EDUCATION -2.413e+02 2.697e+01  
## CLIENT\_ACTIVITYTYPE -1.696e+01 3.873e+01  
## CLIENT\_TOGETHER\_INCOME 1.081e+00 7.377e-03  
## LOAN\_OUTSTANDINGLOANSCOUNT -7.800e+01 5.288e+01  
## LOAN\_EXISTINGCUSTOMERFLAG -6.238e+02 2.092e+02  
## LOAN\_OVERDUE\_EXIST\_FLAG 5.861e+02 1.649e+02  
## SCR\_WORKPLACECONFIRMED 2.103e+03 3.405e+02  
## DECADE -5.112e+01 7.124e+01  
## LOAN\_TERM 1.327e+00 2.135e-01  
## ZODIAC 2.095e+01 1.713e+01  
## ZODIAC\_CHINA -1.819e+01 1.720e+01  
## AGE -7.049e+01 1.165e+01  
## CHANGE\_WORK -2.987e+02 1.981e+02  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE 5.191e+02 1.648e+02  
## FAMILY\_PHONE -4.089e+02 1.347e+02  
## REAL\_ESTATE -1.897e+02 1.340e+02  
## CAR -1.845e+03 2.179e+02  
## INFORMATION\_IN\_MGB -2.060e+03 2.190e+02  
## INFORMATION\_IN\_BKI 6.006e+02 2.844e+02  
## STRABIS 4.840e+00 1.544e+02  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 7.840e+03 5.517e+01  
## DELAY -6.127e+02 1.449e+02  
## t value Pr(>|t|)   
## (Intercept) 3.280 0.001042 \*\*   
## LOAN\_PRODUCT\_TYPE 44.941 < 2e-16 \*\*\*  
## BRANCH\_REGION -3.484 0.000495 \*\*\*  
## CLIENT\_GENDER 6.216 5.24e-10 \*\*\*  
## CLIENT\_FAMILYSTATUS -1.240 0.215044   
## CLIENT\_COUNTDEPENDENTS 0.751 0.452422   
## CLIENT\_TOTALEXPERIENCE 3.547 0.000391 \*\*\*  
## CLIENT\_LASTEXPERIENCE 1.673 0.094294 .   
## CLIENT\_EDUCATION -8.947 < 2e-16 \*\*\*  
## CLIENT\_ACTIVITYTYPE -0.438 0.661457   
## CLIENT\_TOGETHER\_INCOME 146.578 < 2e-16 \*\*\*  
## LOAN\_OUTSTANDINGLOANSCOUNT -1.475 0.140233   
## LOAN\_EXISTINGCUSTOMERFLAG -2.982 0.002870 \*\*   
## LOAN\_OVERDUE\_EXIST\_FLAG 3.554 0.000381 \*\*\*  
## SCR\_WORKPLACECONFIRMED 6.176 6.76e-10 \*\*\*  
## DECADE -0.718 0.473051   
## LOAN\_TERM 6.215 5.27e-10 \*\*\*  
## ZODIAC 1.223 0.221198   
## ZODIAC\_CHINA -1.058 0.290004   
## AGE -6.049 1.50e-09 \*\*\*  
## CHANGE\_WORK -1.508 0.131676   
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE 3.151 0.001633 \*\*   
## FAMILY\_PHONE -3.037 0.002396 \*\*   
## REAL\_ESTATE -1.416 0.156697   
## CAR -8.468 < 2e-16 \*\*\*  
## INFORMATION\_IN\_MGB -9.405 < 2e-16 \*\*\*  
## INFORMATION\_IN\_BKI 2.112 0.034705 \*   
## STRABIS 0.031 0.975000   
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 142.108 < 2e-16 \*\*\*  
## DELAY -4.229 2.36e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6932 on 13706 degrees of freedom  
## Multiple R-squared: 0.8552, Adjusted R-squared: 0.8548   
## F-statistic: 2790 on 29 and 13706 DF, p-value: < 2.2e-16

#Висновок: змінні CLIENT\_COUNTDEPENDENTS, CLIENT\_ACTIVITYTYPE, LOAN\_OUTSTANDINGLOANSCOUNT, DECADE, ZODIAC, ZODIAC\_CHINA, CHANGE\_WORK, REAL\_ESTATE, INFORMATION\_IN\_BKI найменш значущі, коефіцієнт детермінації 0,855.

#as p-value, Pr(>|t|) of variable "type" is higher than significance level (5%), let's exclude this variable from the model  
model\_opt <- lm(data = d\_train, 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)   
summary(model\_opt)

##   
## Call:  
## lm(formula = 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,   
## data = d\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59647 -2542 371 2930 32121   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 1.029e+07 2.785e+06  
## LOAN\_PRODUCT\_TYPE 1.833e+03 4.157e+01  
## BRANCH\_REGION -4.777e+01 1.322e+01  
## CLIENT\_GENDER 7.689e+02 1.252e+02  
## CLIENT\_TOTALEXPERIENCE 4.857e+01 1.205e+01  
## CLIENT\_EDUCATION -2.383e+02 2.694e+01  
## CLIENT\_TOGETHER\_INCOME 1.095e+00 7.209e-03  
## LOAN\_EXISTINGCUSTOMERFLAG -2.197e+03 1.374e+02  
## LOAN\_OVERDUE\_EXIST\_FLAG 7.189e+02 1.408e+02  
## SCR\_WORKPLACECONFIRMED 2.158e+03 3.399e+02  
## AGE -6.618e+01 1.148e+01  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE 4.770e+02 1.640e+02  
## FAMILY\_PHONE -3.758e+02 1.348e+02  
## CAR -1.871e+03 2.175e+02  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 8.078e+03 4.720e+01  
## DELAY -7.650e+02 1.408e+02  
## t value Pr(>|t|)   
## (Intercept) 3.696 0.000220 \*\*\*  
## LOAN\_PRODUCT\_TYPE 44.102 < 2e-16 \*\*\*  
## BRANCH\_REGION -3.615 0.000302 \*\*\*  
## CLIENT\_GENDER 6.143 8.32e-10 \*\*\*  
## CLIENT\_TOTALEXPERIENCE 4.029 5.63e-05 \*\*\*  
## CLIENT\_EDUCATION -8.846 < 2e-16 \*\*\*  
## CLIENT\_TOGETHER\_INCOME 151.920 < 2e-16 \*\*\*  
## LOAN\_EXISTINGCUSTOMERFLAG -15.994 < 2e-16 \*\*\*  
## LOAN\_OVERDUE\_EXIST\_FLAG 5.105 3.35e-07 \*\*\*  
## SCR\_WORKPLACECONFIRMED 6.349 2.23e-10 \*\*\*  
## AGE -5.764 8.38e-09 \*\*\*  
## MATCH\_OF\_THE\_REGISTRATION\_PLACE\_WITH\_THE\_ACTUAL\_RESIDENCE 2.908 0.003642 \*\*   
## FAMILY\_PHONE -2.788 0.005313 \*\*   
## CAR -8.604 < 2e-16 \*\*\*  
## PERCENT\_IN\_THE\_LOAN\_AMOUNT 171.133 < 2e-16 \*\*\*  
## DELAY -5.433 5.62e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6965 on 13720 degrees of freedom  
## Multiple R-squared: 0.8536, Adjusted R-squared: 0.8535   
## F-statistic: 5335 on 15 and 13720 DF, p-value: < 2.2e-16

#Висновок: усі змінні значущі, коефіцієнт детермінації трохи зменшився – 0,836.

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

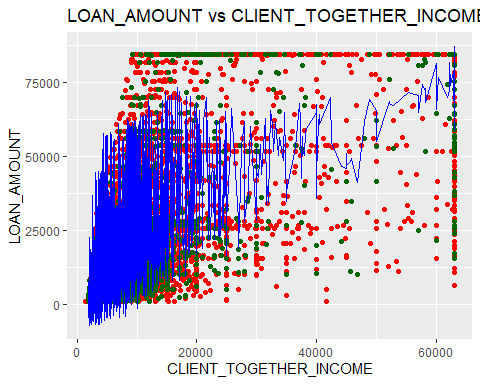
## [1] 48448964

test\_mse\_opt

## [1] 45119952

#Висновок: значення середньоквадратичної помилки покращилися – на навчальній вибірці – 484 489 64, на тестовій вибірці – 451 199 52, тобто перенавчання немає.

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(d\_test$CLIENT\_TOGETHER\_INCOME, p\_mr),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.

d\_train\_poly <- d\_train[,c('LOAN\_AMOUNT', 'CLIENT\_TOGETHER\_INCOME')]  
d\_test\_poly <- d\_test[,c('LOAN\_AMOUNT', 'CLIENT\_TOGETHER\_INCOME')]  
d\_train\_poly$CLIENT\_TOGETHER\_INCOME2 <- d\_train\_poly$CLIENT\_TOGETHER\_INCOME^2  
d\_train\_poly$CLIENT\_TOGETHER\_INCOME3 <- d\_train\_poly$CLIENT\_TOGETHER\_INCOME^3  
d\_test\_poly$CLIENT\_TOGETHER\_INCOME2 <- d\_test\_poly$CLIENT\_TOGETHER\_INCOME^2  
d\_test\_poly$CLIENT\_TOGETHER\_INCOME3 <- d\_test\_poly$CLIENT\_TOGETHER\_INCOME^3  
model\_pr <- lm(data = d\_train\_poly, LOAN\_AMOUNT ~ CLIENT\_TOGETHER\_INCOME2 + CLIENT\_TOGETHER\_INCOME3)  
summary(model\_pr)

##   
## Call:  
## lm(formula = LOAN\_AMOUNT ~ CLIENT\_TOGETHER\_INCOME2 + CLIENT\_TOGETHER\_INCOME3,   
## data = d\_train\_poly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -71218 -8238 -2881 6328 66842   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.183e+04 1.604e+02 73.73 <2e-16 \*\*\*  
## CLIENT\_TOGETHER\_INCOME2 7.822e-05 1.263e-06 61.94 <2e-16 \*\*\*  
## CLIENT\_TOGETHER\_INCOME3 -1.042e-09 2.072e-11 -50.29 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14240 on 13733 degrees of freedom  
## Multiple R-squared: 0.3873, Adjusted R-squared: 0.3873   
## F-statistic: 4341 on 2 and 13733 DF, p-value: < 2.2e-16

#Висновок: додано змінні m2^2 та m2^3. #Висновок: змінні m2^2 та m2^3 значущі, але коефіцієнт детермінації зменшився – 0,387.

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

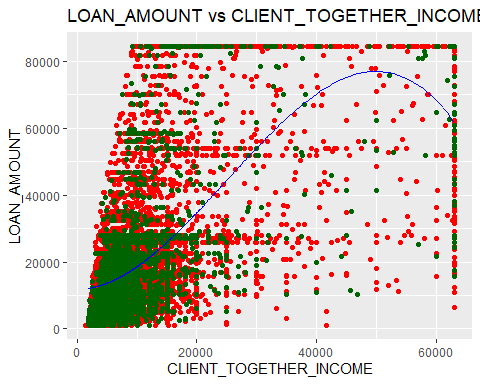
## [1] 202807567

test\_mse\_poly

## [1] 198427351

#Висновок: значення середньоквадратичної помилки трохи зросли на навчальній вибірці – 659 342 728, на тестовій вибірці – 635 313 813, тобто перенавчання немає.

ggplot() +   
 geom\_point(aes(d\_train\_poly$CLIENT\_TOGETHER\_INCOME, d\_train\_poly$LOAN\_AMOUNT), colour = 'red') + geom\_point(aes(d\_test\_poly$CLIENT\_TOGETHER\_INCOME, d\_test\_poly$LOAN\_AMOUNT), colour = 'dark green') +   
 geom\_line(aes(d\_test\_poly$CLIENT\_TOGETHER\_INCOME, p\_pr),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.

fit <- data.frame(p\_sr, p\_mr, p\_pr)  
write.csv2(fit, file = "LOAN\_AMOUNT.csv")