L5

Kovel

30 РѕРєС‚СЏР±СЂСЏ 2020 Рі

# Download the data

path='C:/education.csv'

set.seed(123)  
f <- read.csv(file = path, header = TRUE, encoding = 'UNICOD')   
head (f)

## school sex age reason guardian traveltime studytime activities internet  
## 1 GP F 18 course mother 2 2 no no  
## 2 GP F 17 course father 1 2 no yes  
## 3 GP F 15 other mother 1 2 no yes  
## 4 GP F 15 home mother 1 3 yes yes  
## 5 GP F 16 home father 1 2 no no  
## 6 GP M 16 reputation mother 1 2 yes yes

## Висновок: для побудови моделі використані дані про освіту.

# Statistics

## Descriptive statistics

library (psych)  
describe(f)

## vars n mean sd median trimmed mad min max range skew  
## school\* 1 395 1.12 0.32 1 1.02 0.00 1 2 1 2.38  
## sex\* 2 395 1.47 0.50 1 1.47 0.00 1 2 1 0.11  
## age 3 395 16.70 1.28 17 16.63 1.48 15 22 7 0.46  
## reason\* 4 395 2.26 1.21 2 2.20 1.48 1 4 3 0.41  
## guardian\* 5 395 1.85 0.54 2 1.84 0.00 1 3 2 -0.11  
## traveltime 6 395 1.45 0.70 1 1.31 0.00 1 4 3 1.59  
## studytime 7 395 2.04 0.84 2 1.96 0.00 1 4 3 0.63  
## activities\* 8 395 1.51 0.50 2 1.51 0.00 1 2 1 -0.04  
## internet\* 9 395 1.83 0.37 2 1.91 0.00 1 2 1 -1.78  
## kurtosis se  
## school\* 3.68 0.02  
## sex\* -1.99 0.03  
## age -0.03 0.06  
## reason\* -1.40 0.06  
## guardian\* 0.15 0.03  
## traveltime 2.27 0.04  
## studytime -0.04 0.04  
## activities\* -2.00 0.03  
## internet\* 1.16 0.02

## Висновок: кількість спостережень – 395, кількість змінних – 9, з них якісних – 6, кількісних – 3. Пропущених значень і викидів немає.

## Factors as numeric

f$school <- as.numeric(as.factor(f$school))-1  
f$sex <- as.numeric(as.factor(f$sex))-1  
f$reason <- as.numeric(as.factor(f$reason))-1  
f$guardian <- as.numeric(as.factor(f$guardian))-1  
f$activities <- as.numeric(as.factor(f$activities))-1  
f$internet <- as.numeric(as.factor(f$internet))-1  
head (f)

## school sex age reason guardian traveltime studytime activities internet  
## 1 0 0 18 0 1 2 2 0 0  
## 2 0 0 17 0 0 1 2 0 1  
## 3 0 0 15 2 1 1 2 0 1  
## 4 0 0 15 1 1 1 3 1 1  
## 5 0 0 16 1 0 1 2 0 0  
## 6 0 1 16 3 1 1 2 1 1

## Висновок:моделі класифікації вимагають попереднього шкалювання кількісних змінних.

# Splitting the scaled dataset into the TRAIN set and TEST set

set.seed(123)  
library(caTools)  
split = sample.split(f$school, SplitRatio = 2/3)  
f\_train = subset(f, split == TRUE)  
f\_test = subset(f, split == FALSE)

## Висновок: підготований датасет розділено на навчальну та тестову вибірки.

# Features Scaling

mage <- mean(f\_train$age)  
sage <- sd(f\_train$age)  
mtraveltime <- mean(f\_train$traveltime)  
straveltime <- sd(f\_train$traveltime)  
mstudytime <- mean(f\_train$studytime)  
sstudytime <- sd(f\_train$studytime)  
  
f\_train$age <- (f\_train$age-mage)/sage  
f\_test$age <- (f\_test$age-mage)/sage  
  
f\_train$traveltime <- (f\_train$traveltime-mtraveltime)/straveltime  
f\_test$traveltime <- (f\_test$traveltime-mtraveltime)/straveltime  
  
f\_train$studytime <- (f\_train$studytime-mstudytime)/sstudytime  
f\_test$studytime <- (f\_test$studytime-mstudytime)/sstudytime  
  
head (f\_train)

## school sex age reason guardian traveltime studytime activities  
## 1 0 0 1.0834932 0 1 0.7771603 -0.04947419 0  
## 3 0 0 -1.3823879 2 1 -0.6378075 -0.04947419 0  
## 6 0 1 -0.5604275 3 1 -0.6378075 -0.04947419 1  
## 7 0 1 -0.5604275 1 1 -0.6378075 -0.04947419 0  
## 9 0 1 -1.3823879 1 1 -0.6378075 -0.04947419 0  
## 10 0 1 -1.3823879 1 1 -0.6378075 -0.04947419 1  
## internet  
## 1 0  
## 3 1  
## 6 1  
## 7 1  
## 9 1  
## 10 1

head(f\_test)

## school sex age reason guardian traveltime studytime activities  
## 2 0 0 0.2615328 0 0 -0.6378075 -0.04947419 0  
## 4 0 0 -1.3823879 1 1 -0.6378075 1.13790641 1  
## 5 0 0 -0.5604275 1 0 -0.6378075 -0.04947419 0  
## 8 0 0 0.2615328 1 1 0.7771603 -0.04947419 0  
## 11 0 0 -1.3823879 3 1 -0.6378075 -0.04947419 0  
## 13 0 1 -1.3823879 0 0 -0.6378075 -1.23685479 1  
## internet  
## 2 1  
## 4 1  
## 5 0  
## 8 0  
## 11 1  
## 13 1

## Висновок:моделі класифікації вимагають попереднього шкалювання кількісних змінних.

# Fitting (Benchmark model)

class\_lr <- glm(school ~ ., f\_train, family = binomial)  
summary(class\_lr)

##   
## Call:  
## glm(formula = school ~ ., family = binomial, data = f\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6505 -0.4048 -0.1601 -0.0718 2.5305   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.4111 0.7396 -1.908 0.05639 .   
## sex -0.9534 0.5636 -1.692 0.09071 .   
## age 1.9906 0.3982 4.999 5.77e-07 \*\*\*  
## reason -0.3025 0.2145 -1.411 0.15837   
## guardian -1.4180 0.4395 -3.226 0.00125 \*\*   
## traveltime 0.6621 0.2156 3.071 0.00213 \*\*   
## studytime -0.5328 0.3104 -1.717 0.08601 .   
## activities -0.3719 0.4724 -0.787 0.43105   
## internet 0.2354 0.6072 0.388 0.69827   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 191.01 on 263 degrees of freedom  
## Residual deviance: 123.17 on 255 degrees of freedom  
## AIC: 141.17  
##   
## Number of Fisher Scoring iterations: 7

## Висновок: значущими змінними є age, guardian та traveltime.

## Optimized model

class\_opt <- glm(school ~ age + guardian + traveltime, f\_train, family = binomial)  
summary(class\_opt)

##   
## Call:  
## glm(formula = school ~ age + guardian + traveltime, family = binomial,   
## data = f\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5330 -0.4599 -0.2063 -0.1053 2.6360   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.1731 0.3973 -5.470 4.49e-08 \*\*\*  
## age 1.8242 0.3502 5.209 1.90e-07 \*\*\*  
## guardian -1.2495 0.4008 -3.117 0.00183 \*\*   
## traveltime 0.7797 0.1991 3.917 8.98e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 191.01 on 263 degrees of freedom  
## Residual deviance: 131.27 on 260 degrees of freedom  
## AIC: 139.27  
##   
## Number of Fisher Scoring iterations: 7

## Висновок: всі змінні оптимізованої моделі є значущими, AIC став нижчим

# Predicting

p <- predict(class\_opt, f\_test[, c('age','guardian','traveltime')], 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

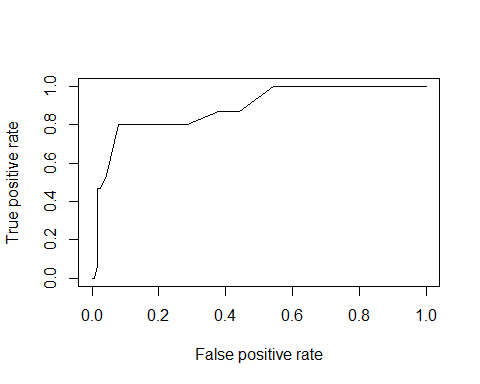
cm = table(f\_test[, 'school'], y > 0.5)  
print(cm)

##   
## FALSE TRUE  
## 0 114 2  
## 1 9 6

## Висновок: точність моделі - (114 + 6) / 131 = 91,6 %, частка невірно класифікованих випадків – (9 + 2) / 131 = 8,3 %. Чутливість моделі – 6 / (9 + 6) = 40 %, специфічність – 114 / (114 + 2) = 98,2 %, тобто модель більш чутлива до виявлення негативних випадків.

## ROC

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

 ##Висновок: співвідношення істинно-позитивних і хибно-позитивних випадків свідчить про досить високу якість моделі. ## Optimized model2

class\_opt2 <- glm(school ~ age + traveltime, f\_train, family = binomial)  
summary(class\_opt2)

##   
## Call:  
## glm(formula = school ~ age + traveltime, family = binomial, data = f\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7508 -0.5198 -0.1860 -0.1072 2.4464   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.8395 0.3325 -8.540 < 2e-16 \*\*\*  
## age 1.3479 0.2658 5.070 3.97e-07 \*\*\*  
## traveltime 0.7119 0.1891 3.764 0.000167 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 191.01 on 263 degrees of freedom  
## Residual deviance: 141.80 on 261 degrees of freedom  
## AIC: 147.8  
##   
## Number of Fisher Scoring iterations: 6

## Висновок: всі змінні оптимізованої моделі є дуже значущими, AIC став трохи вищим

# Predicting

p <- predict(class\_opt2, f\_test[, c('age','traveltime')], 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

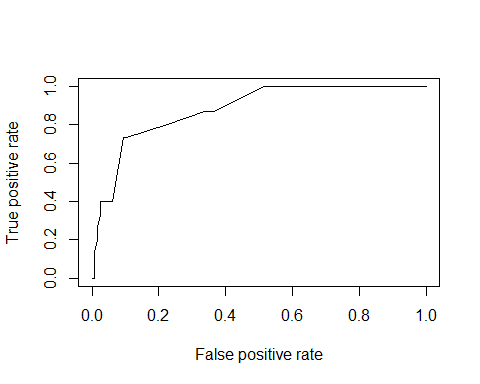
cm = table(f\_test[, 'school'], y > 0.5)  
print(cm)

##   
## FALSE TRUE  
## 0 113 3  
## 1 9 6

## Висновок: точність моделі - (113 + 6) / 131 = 90,8 %, частка невірно класифікованих випадків – (9 + 3) / 131 = 9,1 %. Чутливість моделі – 6 / (9 + 6) = 40 %, специфічність – 113 / (113 + 3) = 97,4 %, тобто модель більш чутлива до виявлення негативних випадків.

## ROC

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

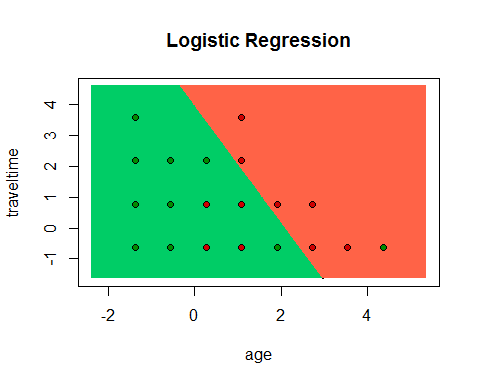
 ##Висновок: співвідношення істинно-позитивних і хибно-позитивних випадків свідчить про досить високу якість моделі. # Visualising the Test set results

library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

set = f\_test[,c('age','traveltime','school')]  
X1 = seq(min(set['age']) - 1, max(set['age']) + 1, by = 0.01)  
X2 = seq(min(set['traveltime']) - 1, max(set['traveltime']) + 1, by = 0.01)  
grid\_set = expand.grid(X1, X2)  
colnames(grid\_set) = c('age', 'traveltime')  
prob\_set = predict(class\_opt2, grid\_set, type = 'response')  
y\_grid = ifelse(prob\_set > 0.5, 1, 0)  
plot(set[, -3],  
 main = 'Logistic Regression',  
 xlab = 'age', ylab = 'traveltime',  
 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, 'tomato', 'springgreen3'))  
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'red3', 'green4'))

 ## Построєно графік. Модель описує лінійний варіант розділяючої кривої. # Write prepared data to the file

write.csv2(f\_train, file = "education\_train.csv")  
write.csv2(f\_test, file = "education\_test.csv")