Entrega2

Numeric and Binary targets Forecasting Models

Pol Renau Miguel Angel Merino

Table of Contents

install.packages(c("car", "rgl"))

## Installing packages into '/home/pol/R/x86\_64-pc-linux-gnu-library/3.6'  
## (as 'lib' is unspecified)

install.packages("cran")

## Installing package into '/home/pol/R/x86\_64-pc-linux-gnu-library/3.6'  
## (as 'lib' is unspecified)

## Warning: package 'cran' is not available (for R version 3.6.1)

library(car)

## Loading required package: carData

library(rgl)  
library(effects)

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

install.packages(c("ROCR"))

## Installing package into '/home/pol/R/x86\_64-pc-linux-gnu-library/3.6'  
## (as 'lib' is unspecified)

library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

install.packages(("FactoMineR"))

## Installing package into '/home/pol/R/x86\_64-pc-linux-gnu-library/3.6'  
## (as 'lib' is unspecified)

library(FactoMineR)

load("mostra2.RData")

# Binary/Logistic Regression Models with Binary target

# Split dataframe in test and work dataframes  
set.seed(14121997)  
ll<-sample(1:nrow(df),nrow(df)\*0.80)  
ll<-order(ll)  
dfwork<-df[ll,]  
dftest<-df[-ll,]

names(df)

## [1] "age" "type.employer" "fnlwgt" "education"   
## [5] "education.num" "marital" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hr.per.week" "country" "y.bin" "f.type"   
## [17] "f.marital" "f.education" "f.continent" "f.benefici"   
## [21] "f.age" "f.hpw" "f.educationNum" "i.rank"

vars\_exp<-c("age","education.num","capital.gain","capital.loss","hr.per.week")  
  
vars\_edisc<-names(dfwork)[c(7:10,16:19,22)]; vars\_edisc

## [1] "occupation" "relationship" "race" "sex"   
## [5] "f.type" "f.marital" "f.education" "f.continent"   
## [9] "f.hpw"

catdes(dfwork[,c("y.bin",vars\_exp)],1)

##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## education.num 0.12007133 1.120696e-107  
## capital.gain 0.08494083 2.624274e-75  
## age 0.05423366 5.551358e-48  
## hr.per.week 0.05324202 4.100827e-47  
## capital.loss 0.02820879 1.896187e-25  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $`<=50K`  
## v.test Mean in category Overall mean sd in category  
## capital.loss -10.35343 51.739340 90.15969 312.752884  
## hr.per.week -14.22391 38.513755 39.75717 9.848726  
## age -14.35576 36.612792 38.36254 13.835669  
## capital.gain -17.96594 153.332659 554.09296 890.818949  
## education.num -21.36050 9.555365 10.04920 2.424236  
## Overall sd p.value  
## capital.loss 412.800862 4.037766e-25  
## hr.per.week 9.724336 6.510879e-46  
## age 13.558495 9.804119e-47  
## capital.gain 2481.406480 3.600981e-72  
## education.num 2.571769 3.114438e-101  
##   
## $`>50K`  
## v.test Mean in category Overall mean sd in category  
## education.num 21.36050 11.65733 10.04920 2.373601  
## capital.gain 17.96594 1859.14441 554.09296 4625.814262  
## age 14.35576 44.06047 38.36254 10.801411  
## hr.per.week 14.22391 43.80627 39.75717 8.074921  
## capital.loss 10.35343 215.27324 90.15969 621.554304  
## Overall sd p.value  
## education.num 2.571769 3.114438e-101  
## capital.gain 2481.406480 3.600981e-72  
## age 13.558495 9.804119e-47  
## hr.per.week 9.724336 6.510879e-46  
## capital.loss 412.800862 4.037766e-25

Segons el catdes obtingut, veiem que totes les variables tenen un p valor inferior a 0.05, per tant seguint el criteri de catdes, deixem totes les variables com a rellevants per començar a modelar. També podem observar que les més rellevants són education.num i darrerament la segueix capital gain.

## Construncció d’un primer model (només numèric)

m1<-glm(y.bin~.,family =binomial,data=dfwork[,c("y.bin",vars\_exp)] )  
summary(m1)

##   
## Call:  
## glm(formula = y.bin ~ ., family = binomial, data = dfwork[, c("y.bin",   
## vars\_exp)])  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3355 -0.6439 -0.4037 -0.1269 2.6971   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.029e+00 3.684e-01 -24.510 < 2e-16 \*\*\*  
## age 4.331e-02 3.627e-03 11.940 < 2e-16 \*\*\*  
## education.num 3.413e-01 2.029e-02 16.815 < 2e-16 \*\*\*  
## capital.gain 3.049e-04 2.872e-05 10.619 < 2e-16 \*\*\*  
## capital.loss 7.118e-04 9.482e-05 7.506 6.09e-14 \*\*\*  
## hr.per.week 5.384e-02 5.401e-03 9.969 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4144.4 on 3800 degrees of freedom  
## Residual deviance: 3070.5 on 3795 degrees of freedom  
## AIC: 3082.5  
##   
## Number of Fisher Scoring iterations: 5

Anova(m1,test="LR") # afegim el LR (CHIsq) ja que es un model generalitzat

## Analysis of Deviance Table (Type II tests)  
##   
## Response: y.bin  
## LR Chisq Df Pr(>Chisq)   
## age 148.54 1 < 2e-16 \*\*\*  
## education.num 334.08 1 < 2e-16 \*\*\*  
## capital.gain 201.63 1 < 2e-16 \*\*\*  
## capital.loss 56.77 1 4.9e-14 \*\*\*  
## hr.per.week 108.95 1 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

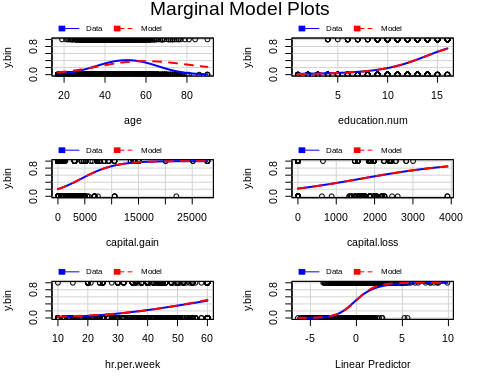
#totes es variables tenen uns efectes nets significatius  
m2<-step(m1, k = log(nrow(dfwork))) # PEr el bBIC utilitzem la K, usant el log

## Start: AIC=3119.91  
## y.bin ~ age + education.num + capital.gain + capital.loss + hr.per.week  
##   
## Df Deviance AIC  
## <none> 3070.5 3119.9  
## - capital.loss 1 3127.2 3168.4  
## - hr.per.week 1 3179.4 3220.6  
## - age 1 3219.0 3260.2  
## - capital.gain 1 3272.1 3313.3  
## - education.num 1 3404.5 3445.8

#Veiem que totes són rellavants

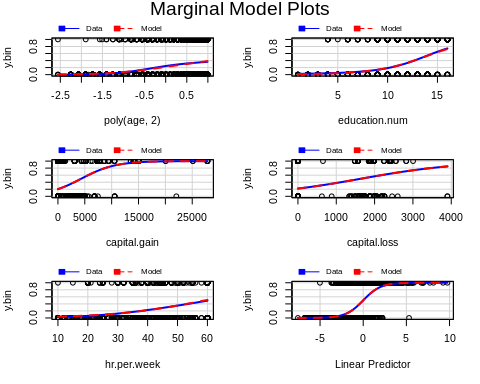
Tot i que amb el catdes ja haviem vist que totes eren rellevants, hem fet un model que usi totes les variables numèriques, i també hem confirmat que totes són necesàries.

marginalModelPlots(m1)

 Com podem apreciar en el marginal model plots, totes les variables numèriques s’adapten de forma adeqüada a excepció de la variable explicativa edat. S’intueix una forma polinomica quadrada, per tant ho probarem a continuació.

m2<- glm(y.bin~poly(age,2)+education.num+capital.gain+capital.loss+hr.per.week,family = binomial, data = dfwork)  
marginalModelPlots(m2)

## Warning in mmps(...): Splines and/or polynomials replaced by a fitted  
## linear combination

 Veiem amb els plots marginals, que hi ha una millor adaptació d’aquesta manera. A continuació comprobarem quin dels dos models que hem calculat de moment és millor.

anova(m1,m2,test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: y.bin ~ age + education.num + capital.gain + capital.loss + hr.per.week  
## Model 2: y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 3795 3070.4   
## 2 3794 2976.5 1 93.92 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Com que el p-valor es inferior a 0.05, rebutjem la hipotesis nul·la, els dos models no són equivalents.

BIC(m1,m2)

## df BIC  
## m1 6 3119.912  
## m2 7 3034.235

Ens quedem amb el model dos ja que com hem vist anteriorment és millor model que el inicial, ja que m2 te un BIC inferior.

## Afegir factors

out<-catdes(dfwork[,c("y.bin",vars\_edisc)],1); out[1]

## $test.chi2  
## p.value df  
## relationship 1.130244e-165 5  
## f.marital 2.067667e-149 3  
## occupation 1.815894e-93 12  
## f.hpw 3.957874e-49 4  
## f.education 8.539953e-35 3  
## sex 2.585052e-29 1  
## f.type 1.940759e-20 3  
## race 5.390885e-09 4

Amb les variables factors, també veiem que ens pasa algo similar que amb les numériques, en principi són rellevants totes a excepció de f.continent, també observem que aquelles que semblen més rellevants són relationship, f.marital ,occupation i f.hpw . Tot i que ens diu que totes són importants, escollirem aquelles que són de major importància (p-valor, més petit)

m3<- glm(y.bin~(poly(age,2)+education.num+capital.gain+capital.loss+hr.per.week)+occupation+relationship+race+sex+f.type+f.marital+f.education+f.hpw,family=binomial,data=dfwork)  
  
anova(m2,m3,test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week  
## Model 2: y.bin ~ (poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week) + occupation + relationship + race + sex + f.type +   
## f.marital + f.education + f.hpw  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 3794 2976.5   
## 2 3759 2338.9 35 637.61 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

BIC(m2,m3)

## df BIC  
## m2 7 3034.235  
## m3 42 2685.132

Un cop construit el model amb els factors afegits, fem un step, per veure quins dels factors afegits no són rellevants Rebutjem hipotesis nul·la de que m2 i m3 són equivalents, i veiem que el model m3 és millor que el model m2, per tant seguim treballant amb el m3.

m4<-step(m3, k = log(nrow(dfwork)))

## Start: AIC=2685.13  
## y.bin ~ (poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week) + occupation + relationship + race + sex + f.type +   
## f.marital + f.education + f.hpw  
##   
## Df Deviance AIC  
## - occupation 12 2400.5 2647.8  
## - race 4 2348.1 2661.3  
## - f.education 3 2340.6 2662.1  
## - f.hpw 4 2351.3 2664.5  
## - f.marital 3 2343.9 2665.4  
## - f.type 3 2352.6 2674.1  
## - hr.per.week 1 2339.8 2677.8  
## <none> 2338.9 2685.1  
## - sex 1 2350.5 2688.4  
## - capital.loss 1 2374.0 2712.0  
## - relationship 5 2407.1 2712.1  
## - poly(age, 2) 2 2408.5 2738.2  
## - education.num 1 2444.4 2782.4  
## - capital.gain 1 2503.7 2841.6  
##   
## Step: AIC=2647.75  
## y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week + relationship + race + sex + f.type + f.marital +   
## f.education + f.hpw  
##   
## Df Deviance AIC  
## - race 4 2410.3 2624.7  
## - f.hpw 4 2412.0 2626.3  
## - f.education 3 2404.2 2626.7  
## - f.marital 3 2404.6 2627.1  
## - f.type 3 2415.4 2638.0  
## - hr.per.week 1 2400.8 2639.8  
## <none> 2400.5 2647.8  
## - sex 1 2411.7 2650.7  
## - capital.loss 1 2437.5 2676.6  
## - relationship 5 2472.0 2678.1  
## - poly(age, 2) 2 2477.8 2708.6  
## - capital.gain 1 2574.0 2813.0  
## - education.num 1 2631.3 2870.3  
##   
## Step: AIC=2624.65  
## y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week + relationship + sex + f.type + f.marital + f.education +   
## f.hpw  
##   
## Df Deviance AIC  
## - f.hpw 4 2421.7 2603.0  
## - f.education 3 2413.8 2603.4  
## - f.marital 3 2414.1 2603.7  
## - f.type 3 2425.7 2615.2  
## - hr.per.week 1 2410.7 2616.8  
## <none> 2410.3 2624.7  
## - sex 1 2421.5 2627.6  
## - capital.loss 1 2448.8 2654.9  
## - relationship 5 2483.9 2657.0  
## - poly(age, 2) 2 2488.7 2686.5  
## - capital.gain 1 2582.1 2788.1  
## - education.num 1 2644.2 2850.2  
##   
## Step: AIC=2603.04  
## y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week + relationship + sex + f.type + f.marital + f.education  
##   
## Df Deviance AIC  
## - f.education 3 2425.3 2581.9  
## - f.marital 3 2425.3 2581.9  
## - f.type 3 2437.2 2593.9  
## <none> 2421.7 2603.0  
## - sex 1 2433.4 2606.6  
## - hr.per.week 1 2453.4 2626.5  
## - capital.loss 1 2461.9 2635.0  
## - relationship 5 2495.8 2635.9  
## - poly(age, 2) 2 2499.7 2664.5  
## - capital.gain 1 2595.3 2768.4  
## - education.num 1 2656.4 2829.5  
##   
## Step: AIC=2581.91  
## y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week + relationship + sex + f.type + f.marital  
##   
## Df Deviance AIC  
## - f.marital 3 2428.9 2560.8  
## - f.type 3 2440.8 2572.7  
## <none> 2425.3 2581.9  
## - sex 1 2437.0 2585.3  
## - hr.per.week 1 2456.5 2604.9  
## - relationship 5 2498.9 2614.3  
## - capital.loss 1 2466.2 2614.6  
## - poly(age, 2) 2 2502.8 2643.0  
## - capital.gain 1 2598.3 2746.7  
## - education.num 1 2759.5 2907.9  
##   
## Step: AIC=2560.81  
## y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week + relationship + sex + f.type  
##   
## Df Deviance AIC  
## - f.type 3 2444.3 2551.5  
## <none> 2428.9 2560.8  
## - sex 1 2439.5 2563.2  
## - hr.per.week 1 2458.9 2582.5  
## - capital.loss 1 2469.4 2593.1  
## - poly(age, 2) 2 2513.3 2628.7  
## - capital.gain 1 2602.7 2726.3  
## - education.num 1 2763.2 2886.8  
## - relationship 5 2880.1 2970.8  
##   
## Step: AIC=2551.51  
## y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +   
## hr.per.week + relationship + sex  
##   
## Df Deviance AIC  
## <none> 2444.3 2551.5  
## - sex 1 2454.9 2553.8  
## - hr.per.week 1 2476.9 2575.8  
## - capital.loss 1 2484.5 2583.4  
## - poly(age, 2) 2 2530.3 2621.0  
## - capital.gain 1 2621.2 2720.1  
## - education.num 1 2791.9 2890.8  
## - relationship 5 2899.1 2965.1

summary(m4)

##   
## Call:  
## glm(formula = y.bin ~ poly(age, 2) + education.num + capital.gain +   
## capital.loss + hr.per.week + relationship + sex, family = binomial,   
## data = dfwork)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.5822 -0.5085 -0.1961 -0.0289 3.4345   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.187e+00 4.368e-01 -16.453 < 2e-16 \*\*\*  
## poly(age, 2)1 3.466e+01 4.174e+00 8.303 < 2e-16 \*\*\*  
## poly(age, 2)2 -2.485e+01 4.224e+00 -5.883 4.02e-09 \*\*\*  
## education.num 3.927e-01 2.367e-02 16.590 < 2e-16 \*\*\*  
## capital.gain 2.974e-04 2.971e-05 10.008 < 2e-16 \*\*\*  
## capital.loss 6.796e-04 1.091e-04 6.230 4.67e-10 \*\*\*  
## hr.per.week 3.645e-02 6.492e-03 5.615 1.96e-08 \*\*\*  
## relationshipNot-in-family -1.744e+00 1.596e-01 -10.926 < 2e-16 \*\*\*  
## relationshipOther-relative -2.243e+00 6.086e-01 -3.685 0.000229 \*\*\*  
## relationshipOwn-child -3.147e+00 4.717e-01 -6.672 2.53e-11 \*\*\*  
## relationshipUnmarried -2.063e+00 2.990e-01 -6.899 5.25e-12 \*\*\*  
## relationshipWife 1.430e+00 2.810e-01 5.090 3.58e-07 \*\*\*  
## sexMale 6.853e-01 2.137e-01 3.206 0.001345 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4144.4 on 3800 degrees of freedom  
## Residual deviance: 2444.4 on 3788 degrees of freedom  
## AIC: 2470.4  
##   
## Number of Fisher Scoring iterations: 7

Ens quedem com a factors rellevants amb “sex” i “relationship”

## Afegir interaccions

m5<- glm(y.bin~(poly(age,2)+education.num+capital.gain+capital.loss+hr.per.week)\*(sex+relationship), family = binomial, data = dfwork)

Un cop creat el model amb interaccions, anem a veure si el podem reduïr, eliminant coses irrellevants.

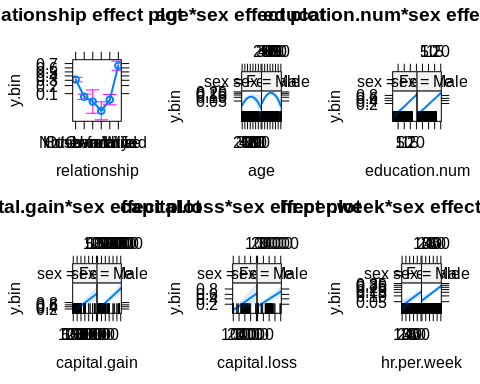
m5a<- glm(y.bin~(poly(age,2)+education.num+capital.gain+capital.loss+hr.per.week)\*(sex)+relationship, family = binomial, data = dfwork)  
m5b<- glm(y.bin~(poly(age,2)+education.num+capital.gain+capital.loss+hr.per.week)\*(relationship)+sex, family = binomial, data = dfwork)  
  
BIC(m5,m5a,m5b)

## df BIC  
## m5 49 2812.123  
## m5a 19 2599.414  
## m5b 43 2765.169

Veiem que afegint les intereccions, el millor model es el que només afegeix interacció amb sex. Per tant deixarem la interacció de sex amb la resta de variables

## Detecció outliers i influents

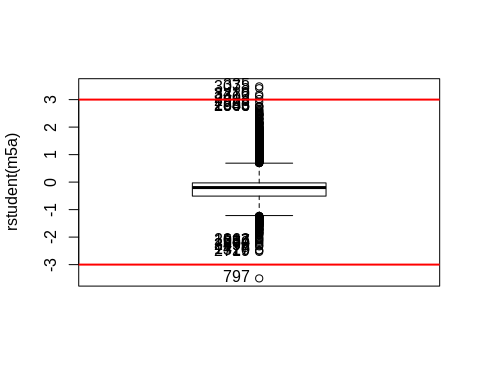
plot(allEffects(m5a))



Boxplot(rstudent(m5a), id.n=2)

## [1] 797 2719 320 1477 2590 3492 679 984 1537 2063 335 3078 3420 3312  
## [15] 2403 2933 885 2386 2646 1903

abline(h=c(3,-3),col="red",lwd=2)

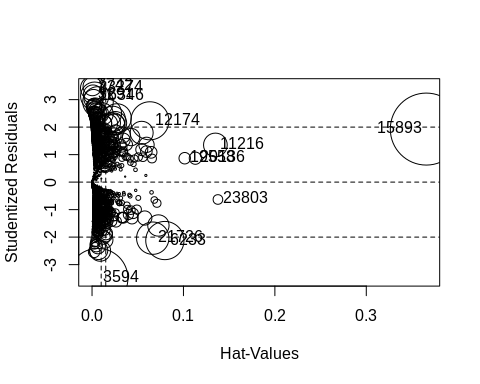


out <- which(rstudent(m5a) >= 3 | rstudent(m5a) <= -3);length(out)

## [1] 6

Podem apreciar que hi han poc outliers, i no són molt extrems per tant no els tindrem en compte. Per cert cal remarcar que hem apujat el valor a 3, per intentar no ser massa estrictes.

influencePlot(m5a,id=list(method="noteworthy",n=5))



## StudRes Hat CookD  
## 7742 3.4804430 0.0003549003 0.007419543  
## 21726 -2.0379463 0.0664106124 0.022251566  
## 3594 -3.5008811 0.0064613999 0.076438448  
## 20136 0.8701431 0.1130183110 0.003186946  
## 19558 0.8690220 0.1013903651 0.002803819  
## 15893 1.9286820 0.3656611449 0.110253218  
## 6233 -2.1242742 0.0796534820 0.031064980  
## 12174 2.2294956 0.0632544482 0.030825607  
## 23803 -0.6314765 0.1377087571 0.001961166  
## 23274 3.4141498 0.0008187439 0.012927775  
## 16346 3.1271605 0.0028474362 0.016914158  
## 3291 3.1852718 0.0010063356 0.007814177  
## 11216 1.3513376 0.1349288350 0.011946460

## Model Final

mfinal<-m5a

### Interpretació del model

summary(mfinal)

##   
## Call:  
## glm(formula = y.bin ~ (poly(age, 2) + education.num + capital.gain +   
## capital.loss + hr.per.week) \* (sex) + relationship, family = binomial,   
## data = dfwork)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2883 -0.5047 -0.1974 -0.0292 3.4601   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.177e+00 8.552e-01 -8.392 < 2e-16 \*\*\*  
## poly(age, 2)1 2.915e+01 8.727e+00 3.340 0.000839 \*\*\*  
## poly(age, 2)2 -2.209e+01 8.178e+00 -2.701 0.006910 \*\*   
## education.num 4.012e-01 5.630e-02 7.126 1.03e-12 \*\*\*  
## capital.gain 2.562e-04 5.317e-05 4.818 1.45e-06 \*\*\*  
## capital.loss 7.462e-04 2.382e-04 3.133 0.001732 \*\*   
## hr.per.week 3.518e-02 1.281e-02 2.745 0.006046 \*\*   
## sexMale 6.498e-01 9.638e-01 0.674 0.500171   
## relationshipNot-in-family -1.739e+00 1.607e-01 -10.827 < 2e-16 \*\*\*  
## relationshipOther-relative -2.231e+00 6.114e-01 -3.649 0.000263 \*\*\*  
## relationshipOwn-child -3.169e+00 4.741e-01 -6.684 2.33e-11 \*\*\*  
## relationshipUnmarried -2.037e+00 3.018e-01 -6.750 1.48e-11 \*\*\*  
## relationshipWife 1.405e+00 3.019e-01 4.653 3.27e-06 \*\*\*  
## poly(age, 2)1:sexMale 7.248e+00 9.912e+00 0.731 0.464640   
## poly(age, 2)2:sexMale -4.165e+00 9.546e+00 -0.436 0.662640   
## education.num:sexMale -8.812e-03 6.198e-02 -0.142 0.886945   
## capital.gain:sexMale 5.886e-05 6.414e-05 0.918 0.358786   
## capital.loss:sexMale -8.355e-05 2.679e-04 -0.312 0.755140   
## hr.per.week:sexMale 1.406e-03 1.487e-02 0.095 0.924689   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4144.4 on 3800 degrees of freedom  
## Residual deviance: 2442.8 on 3782 degrees of freedom  
## AIC: 2480.8  
##   
## Number of Fisher Scoring iterations: 7

Farem una predicció manual per la següent entrada: age = 22 education.num = 12 capital.gain = capital.loss = 0 hr.per.week = 25 sexMale Unmarried

predict = -7.18 + 21*(0.29 -0.23)+ 12*0.4 + 25\*(0.035) + 0.65 - 2.04 +7.25 -4.17 -0.009 + 0.00006 -0.00008 + 0.0014

Dona com a resultat 1.43738 que com sabem aixó vol dir que tindrà bastantes possibilitats de guanyar més de 50k anuals.

## Evaluació Del model

### Amb el dfWork

#### Confusion Matrix

Creem la confusion matrix del dfwork

predictModelFinal<- factor(ifelse(predict(mfinal,type="response")<0.5,0,1),labels=c("Pre-<50","pre->50"))  
conf<-table(predictModelFinal,dfwork$y.bin);conf

##   
## predictModelFinal <=50K >50K  
## Pre-<50 2728 388  
## pre->50 180 505

Com bé sabem la diagonal ens indica quan descendent d’esquerra a dreta, es la que ens indica quan bo es el nostre model, quan major és millor es el model.

#### Capacitat predictiva del model

#Capacitat predictiva   
perpc<-100\*(sum(diag(conf))/sum(nrow(dfwork)))  
print(perpc)

## [1] 85.05656

m0<-glm(y.bin~1,family=binomial,data=dfwork) #model nul  
m0<- factor(ifelse(predict(m0,type="response")<0.5,0,1),labels=c("Pre-<50"))  
conf0<-table(m0,dfwork$y.bin);  
#Capacitat predictiva   
perpc0<-100\*(conf0[1,1]/sum(nrow(dfwork)))  
print(perpc0)

## [1] 76.50618

Tenim una capacitat predictiva del 85.05656, el comparem amb el model nul, i veiem que hem conseguit un increment d’un 10 % practicament, amb el que ens quedem conformes amb el resultat obtingut.

### Amb el dftest

En aquest apartat comprobarem que el model no esta sobre ajustat, i que funciona de forma correcte sobre dades que no ha vist al entrenar el model.

#### Confusion Matrix

Creem la confusion matrix del dftest

#taula de confusió  
predictModelFinalTest<- factor(ifelse(predict(mfinal,newdata=dftest,type="response")<0.5,0,1),labels=c("Pre-<50","pre->50"))  
confTest<-table(predictModelFinalTest,dftest$y.bin);confTest

##   
## predictModelFinalTest <=50K >50K  
## Pre-<50 672 85  
## pre->50 50 144

#### Capacitat predictiva

perpcTest<-100\*(sum(diag(confTest))/sum(nrow(dftest)))  
print(perpcTest)

## [1] 85.80442

#Capacitat Predictiva del 85 per cent  
  
m0<-glm(y.bin~1,family=binomial,data=dfwork) #model nul  
m0Test<- factor(ifelse(predict(m0,newdata=dftest,type="response")<0.5,0,1),labels=c("Pre-<50"))  
conf0Test<-table(m0Test,dftest$y.bin);  
#Capacitat predictiva   
perpc0Test<-100\*(conf0Test[1,1]/sum(nrow(dftest)))  
print(perpc0Test)

## [1] 75.92008

Hem vist que tant amb les dades de test com amb les de work, el model ens dona una resposta bastant bona, i acceptable. Veiem que els models no s’havien sobre ajustat.