Examen practica Ejercicio 3

Agustin Lopez Fredes

2023-07-15

Table of Contents

[Ejercicio 3 1](#_Toc140438541)

[1. Ajustar un modelo logístico para predecir la probabilidad de incurrir en mora. 1](#_Toc140438542)

[2. Evaluar la calidad de ajuste del modelo con al menos dos criterios distintos. 5](#_Toc140438543)

[3. Interpretar los coeficientes del modelo elegido. 8](#_Toc140438544)

[4. Evaluar la calidad de clasificacion y compararlo con otro método de clasificación. 9](#_Toc140438545)

## Ejercicio 3

En el archivo morosos.xlsx se encuentran los registros de 10 mil clientes de un banco. Se desea modelar la probabilidad de mora de un cliente en funcion de las variables disponibles en la base.

### 1. Ajustar un modelo logístico para predecir la probabilidad de incurrir en mora.

library(readxl)  
library(ggplot2)  
library(dplyr)

##   
## 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

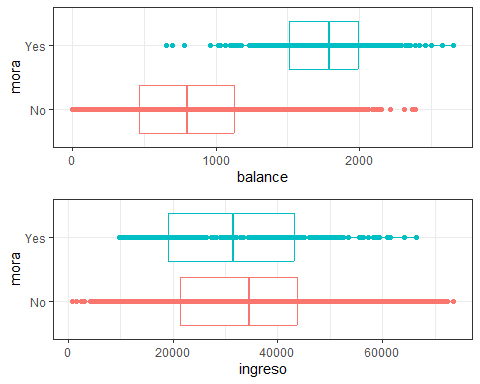
data\_morosos <- read\_excel("morosos.xlsx")  
  
head(data\_morosos)

## # A tibble: 6 × 5  
## orden mora estudiante balance ingreso  
## <dbl> <chr> <chr> <dbl> <dbl>  
## 1 1 No No 730. 44362.  
## 2 2 No Yes 817. 12106.  
## 3 3 No No 1074. 31767.  
## 4 4 No No 529. 35704.  
## 5 5 No No 786. 38463.  
## 6 6 No Yes 920. 7492.

table(data\_morosos$mora)

##   
## No Yes   
## 9667 333

#Grafico los datos  
  
p1 <- ggplot(data=data\_morosos,aes(x = balance, y= mora)) +   
 geom\_boxplot(aes(color = mora)) + geom\_point(aes(color = mora)) +   
 theme\_bw() + theme(legend.position = "null")  
  
p2 <- ggplot(data=data\_morosos,aes(x = ingreso, y= mora)) +   
 geom\_boxplot(aes(color = mora)) + geom\_point(aes(color = mora)) +   
 theme\_bw() + theme(legend.position = "null")  
  
gridExtra::grid.arrange(p1, p2, nrow = 2)



En el boxplot de balance el gráfico indicaría que los clasificados como mora tendrían un balance más alto, con lo que esta variable podría llegar a ser explicativa Los boxplot de ingreso se encuentran bastante solapados, por lo que la diferencia de medianas no parecería ser significativa Por último, podemos observar que hay un gran desbalance de clases, siendo mora=Yes la minoritaria

data\_morosos %>% group\_by(estudiante, mora) %>%   
 summarise(cantidad=n())

## `summarise()` has grouped output by 'estudiante'. You can override using the  
## `.groups` argument.

## # A tibble: 4 × 3  
## # Groups: estudiante [2]  
## estudiante mora cantidad  
## <chr> <chr> <int>  
## 1 No No 6850  
## 2 No Yes 206  
## 3 Yes No 2817  
## 4 Yes Yes 127

En los que no son estudiantes, el 96,48% no está en mora (6850/(6850+206)). En los que sí son estudiantes, ese porcentaje baja a 95,69%. Con el modelo observaremos si esa diferencia contribuye a predecir la clase de “mora”

#Cambiamos los Yes y No de la variable mora por 1s y 0s  
mora1\_nomora0<-ifelse(data\_morosos$mora=="Yes",1,0)  
table(mora1\_nomora0)

## mora1\_nomora0  
## 0 1   
## 9667 333

modelo\_logistico <- glm(mora1\_nomora0 ~ balance+ingreso+estudiante,  
 data=data\_morosos, family = "binomial")   
summary(modelo\_logistico)

##   
## Call:  
## glm(formula = mora1\_nomora0 ~ balance + ingreso + estudiante,   
## family = "binomial", data = data\_morosos)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4691 -0.1418 -0.0557 -0.0203 3.7383   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.087e+01 4.923e-01 -22.080 < 2e-16 \*\*\*  
## balance 5.737e-03 2.319e-04 24.738 < 2e-16 \*\*\*  
## ingreso 3.033e-06 8.203e-06 0.370 0.71152   
## estudianteYes -6.468e-01 2.363e-01 -2.738 0.00619 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2920.6 on 9999 degrees of freedom  
## Residual deviance: 1571.5 on 9996 degrees of freedom  
## AIC: 1579.5  
##   
## Number of Fisher Scoring iterations: 8

El modelo arroja como variables significativas balance y estudiante, ya que el p-valor es menor a 0,05, mientras que ingreso tiene un p-valor de 0,7, indicando que no es significativa

#Generamos los intervalos de confianza del Odds Ratio   
exp(confint(object = modelo\_logistico, level = 0.95))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 7.074481e-06 0.0000487808  
## balance 1.005309e+00 1.0062238757  
## ingreso 9.999870e-01 1.0000191246  
## estudianteYes 3.298827e-01 0.8334223982

Observamos lo mismo, ya que el intervalo de confianza del Odds ratio de ingreso incluye al 1, mientras que balance está por arriba del 1, y estudiantesSi por debajo del 1

Generamos el modelo final con las variables significativas, analizando si la interacción es significativa

modelo\_logistico\_inter <- glm(mora1\_nomora0 ~ balance\*estudiante,  
 data=data\_morosos, family = "binomial")   
summary(modelo\_logistico\_inter)

##   
## Call:  
## glm(formula = mora1\_nomora0 ~ balance \* estudiante, family = "binomial",   
## data = data\_morosos)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4839 -0.1415 -0.0553 -0.0202 3.7628   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.087e+01 4.640e-01 -23.438 <2e-16 \*\*\*  
## balance 5.819e-03 2.937e-04 19.812 <2e-16 \*\*\*  
## estudianteYes -3.512e-01 8.037e-01 -0.437 0.662   
## balance:estudianteYes -2.196e-04 4.781e-04 -0.459 0.646   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2920.6 on 9999 degrees of freedom  
## Residual deviance: 1571.5 on 9996 degrees of freedom  
## AIC: 1579.5  
##   
## Number of Fisher Scoring iterations: 8

Observamos que la interacción no es significativa, por lo que en el modelo final no incluiremos la interacción, ni la variable ingreso descartada previamente, y conservamos las variables balance y estudiante que resultaron significativas

modelo\_logistico\_final <- glm(mora1\_nomora0 ~ balance+estudiante,  
 data=data\_morosos, family = "binomial")   
summary(modelo\_logistico\_final)

##   
## Call:  
## glm(formula = mora1\_nomora0 ~ balance + estudiante, family = "binomial",   
## data = data\_morosos)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4578 -0.1422 -0.0559 -0.0203 3.7435   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.075e+01 3.692e-01 -29.116 < 2e-16 \*\*\*  
## balance 5.738e-03 2.318e-04 24.750 < 2e-16 \*\*\*  
## estudianteYes -7.149e-01 1.475e-01 -4.846 1.26e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2920.6 on 9999 degrees of freedom  
## Residual deviance: 1571.7 on 9997 degrees of freedom  
## AIC: 1577.7  
##   
## Number of Fisher Scoring iterations: 8

Comprobamos que las dos variables son significativas

### 2. Evaluar la calidad de ajuste del modelo con al menos dos criterios distintos.

Utilizando Test de Hosmer- Lemeshow:

#Utilizando Test de Hosmer- Lemeshow  
ResourceSelection::hoslem.test(mora1\_nomora0, fitted(modelo\_logistico\_final))

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: mora1\_nomora0, fitted(modelo\_logistico\_final)  
## X-squared = 3.3586, df = 8, p-value = 0.9099

Observamos que en el Test de Hosmer- Lemeshow no se rechaza la hipótesis nula que sostiene que la distribución de observados y predichos es similar

Evaluando por curva ROC:

# Se obtienen las probabilidades predichas para cada clase   
predicciones <- predict(object = modelo\_logistico\_final, newdata = data\_morosos, type = "response")   
curva\_roc <- pROC::roc(response = mora1\_nomora0, predictor = predicciones)

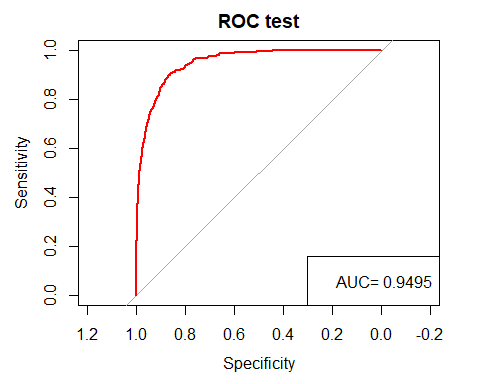
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

curva\_roc

##   
## Call:  
## roc.default(response = mora1\_nomora0, predictor = predicciones)  
##   
## Data: predicciones in 9667 controls (mora1\_nomora0 0) < 333 cases (mora1\_nomora0 1).  
## Area under the curve: 0.9495

# Gráfico de la curva   
#plot(curva\_roc)  
plot(curva\_roc,col="red",lwd=2,main="ROC test")  
legend("bottomright",legend=paste("AUC=",round(pROC::auc(curva\_roc),4)))

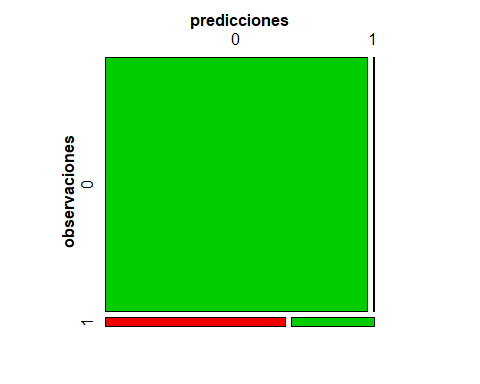


Vemos que el valor de Área Bajo la Curva ROC es superior a 0,5 y cercano a 1, lo que indica una buena calidad de ajuste Debido a que en el punto 1 vimos que hay un gran desbalance de clases, vamos a incorporar también la métrica de recall o sensibilidad, que nos permite evaluar cuántos de los positivos identifica nuestro modelo (siendo los positivos en este caso la clase mayoritaria)

predicciones <- ifelse(test = modelo\_logistico\_final$fitted.values > 0.5, yes = 1, no = 0)   
matriz\_confusion <- table(mora1\_nomora0, predicciones, dnn = c("observaciones", "predicciones"))   
matriz\_confusion

## predicciones  
## observaciones 0 1  
## 0 9628 39  
## 1 228 105

vcd::mosaic(matriz\_confusion, shade = T, colorize = T, gp = grid::gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))



sensibilidad <- matriz\_confusion[4]/(matriz\_confusion[4] + matriz\_confusion[2])  
sensibilidad

## [1] 0.3153153

Observamos que el recall o sensibilidad del modelo es de apenas 31,53%, lo que muestra que el modelo no clasifica correctamente a la clase positiva (mora=Si)

### 3. Interpretar los coeficientes del modelo elegido.

coef\_balance <- modelo\_logistico\_final$coefficients["balance"]  
OR\_balance <- exp(coef\_balance)  
coef\_estudianteSI <- modelo\_logistico\_final$coefficients["estudianteYes"]  
OR\_estudianteSI <- exp(coef\_estudianteSI)  
  
coef\_balance

## balance   
## 0.005738104

OR\_balance

## balance   
## 1.005755

coef\_estudianteSI

## estudianteYes   
## -0.7148776

OR\_estudianteSI

## estudianteYes   
## 0.489252

a) El coeficiente asociado a balance es positivo, lo que nos indica que la probabilidad de mora aumenta con el aumento del balance (saldo al 31/12 pasado). b) El coeficiente de balance es 0.005738104, lo que arroja un Odds Ratio de 1.005755. Esto implica que el odds u oportunidad para balance = X +1 es 1.005755 el odds de balance = x c)El coeficiente asociado a estudianteSi es negativo, lo que indica que la probabilidad de mora disminuye cuando se trata de un estudiante. El Odds Ratio para estudianteSI es de 0.489252, lo que implica que el odds de mora para un estudiante es 0.489 veces el odds de mora en un no estudiante. d)Como el cambio de odds entre los estudiantes y no estudiantes no depende del valor de la otra variable (balance), decimos también “independientemente” del valor de la variable balance

### 4. Evaluar la calidad de clasificacion y compararlo con otro método de clasificación.

Acá tengo duda de qué otro método de clasificación comparar??