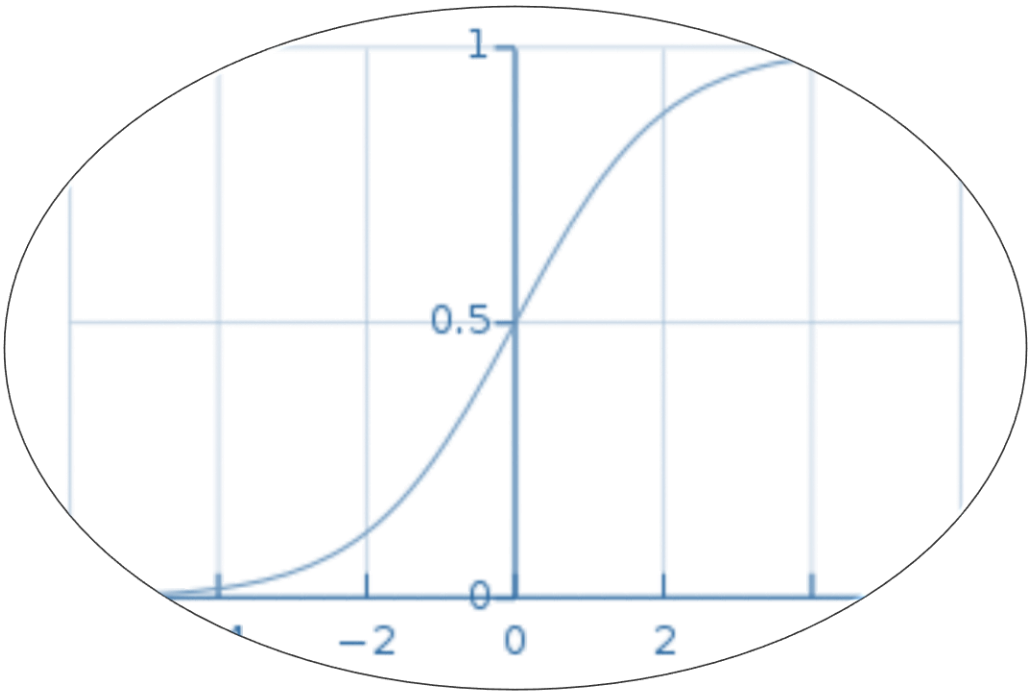
# MASTER BIG DATA

# ESTADÍSTICA: PRÁCTICA 2 (REGRESIÓN LOGÍSITCA)

$$\log\left(\frac{\pi}{1-\pi}\right) = B_0 + B_1 * X_1 + \ldots + B_p * X_p$$




**Autor:** Jaume Feliubadaló Rubio

**Tutor:** Jordi Cortés Martínez

**Data:** 09/12/2018

La Salle, Universidad Ramon Llull

ÍNDICE



ÍNDICE DE FIGURAS

# 1. INTRODUCCIÓN

## 1.1. Resumen

En este trabajo ponemos en práctica el concepto de regresión logística. Siendo este tipo de regresión un tipo de análisis usado con el fin de la predicción del resultado de una variable categórica en función de otras variables independientes o predictoras. Este tipo de análisis se enmarca como ya sabemos en el conjunto de GLM, haciendo uso de la función logit.

(Esta función link (logit en nuestro caso), es una función de conexión 'link', que convierte la probabilidad de éxito entre 0 y 1 en una variable respuesta que tome valores en todos los reales).

## 1.2 Objetivo

El principal objetivo es entender con el concepto explicado ya en clase de regresión logística, prediciendo la probabilidad que un cliente al cual se realiza la llamada acepte el producto que se le ofrece.

$$log\left(\frac{\pi}{1-\pi}\right) = \left(b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p\right)$$

## 1.3 Variables

Cabe destacar que las variables son tanto categóricas como numéricas. Si introducimos variables categóricas en el modelo, R considera que un nivel es el de referencia y le atribuye el valor 0. El resto de niveles se comparando con dicho nivel de referencia. Este concepto se vio en la práctica 1.

# 2. CONSTRUCCIÓN DEL MODELO

## 2.1 Descripción variables

> "*El objetivo de esta práctica es predecir la probabilidad que un cliente al cual se realiza la llamada acepte el producto que se le ofrece*".

Dentro de los datos proporcionados hay las diferentes variables:

Características del cliente

*id*: identificador del cliente
*age*: edad
*job*: tipo de trabajo (admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)
*marital*: estado civil (divorced, married, single, unknown)
*education*: nivel de estudios (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)
*default*: ¿es moroso? (no,yes,unknown)
*housing*: ¿tiene hipoteca? (no,yes,unknown)
*loan*: ¿tiene un préstamo personal? (no,yes,unknown)

Características de la llamada

*contact*: tipo de teléfono (cellular, telephone)
*month*: mes
*day_of_week*: día de la semana (mon, tue, wed, thu, fri)

Otros atributos

*campaign*: número de contactos realizados esta campaña para este cliente (incluyendo el actual)
*pdays*: número de días que han pasado desde que el cliente fue contactado por última vez para una campaña previa (999 significa que no fue contactado previamente)
*previous*: número de llamadas realizadas a este cliente antes de esta campaña
*poutcome*: resultado de la anterior campaña (failure, nonexistent, success)

Indicadores del contexto social y económico

*emp.var.rate*: indicador de la tasa de empleo (cuatrimestral)
*cons.price.idx*: IPC (mensual)
*cons.conf.idx*: Índice de confianza del consumidor (mensual)
*euribor3m*: euribor a 3 meses (diario)
*nr.employed*: número de empleados (cuadrimestral)

Variable respuesta (sólo en el juego de entrenamiento):

*Y*: ¿Se suscribió el cliente al depósito? (yes,no)

**2.2 Modelo_entrenamiento**

Definidas las variables se procede a seguir los pasos indicados en el script guía proporcionado por el tutor.

# Ver Script R con indicaciones en cada etapa del proceso #

Con tal de sintetizar, en este documento solo daré la información que creo necesaria, tal como la interpretación de resultados o las decisiones tomadas bajo criterio propio.

En la figura 1 se aprecia el resultado de la instrucción sapply(datos,class), donde se da el tipo de clase de cada variable en los datos de entrenamiento. Se observa que hay variables de tipo factor, integer y numeric.

```
       age          job      marital    education      default      housing         loan      contact        month  day_of_week
 "integer"     "factor"     "factor"     "factor"     "factor"     "factor"     "factor"     "factor"     "factor"     "factor"
  campaign     previous     poutcome emp.var.rate cons.price.idx cons.conf.idx    euribor3m nr.employed            y
 "integer"    "integer"     "factor"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"     "factor"
```

*Figura 1 Clasificación según tipo de variable*

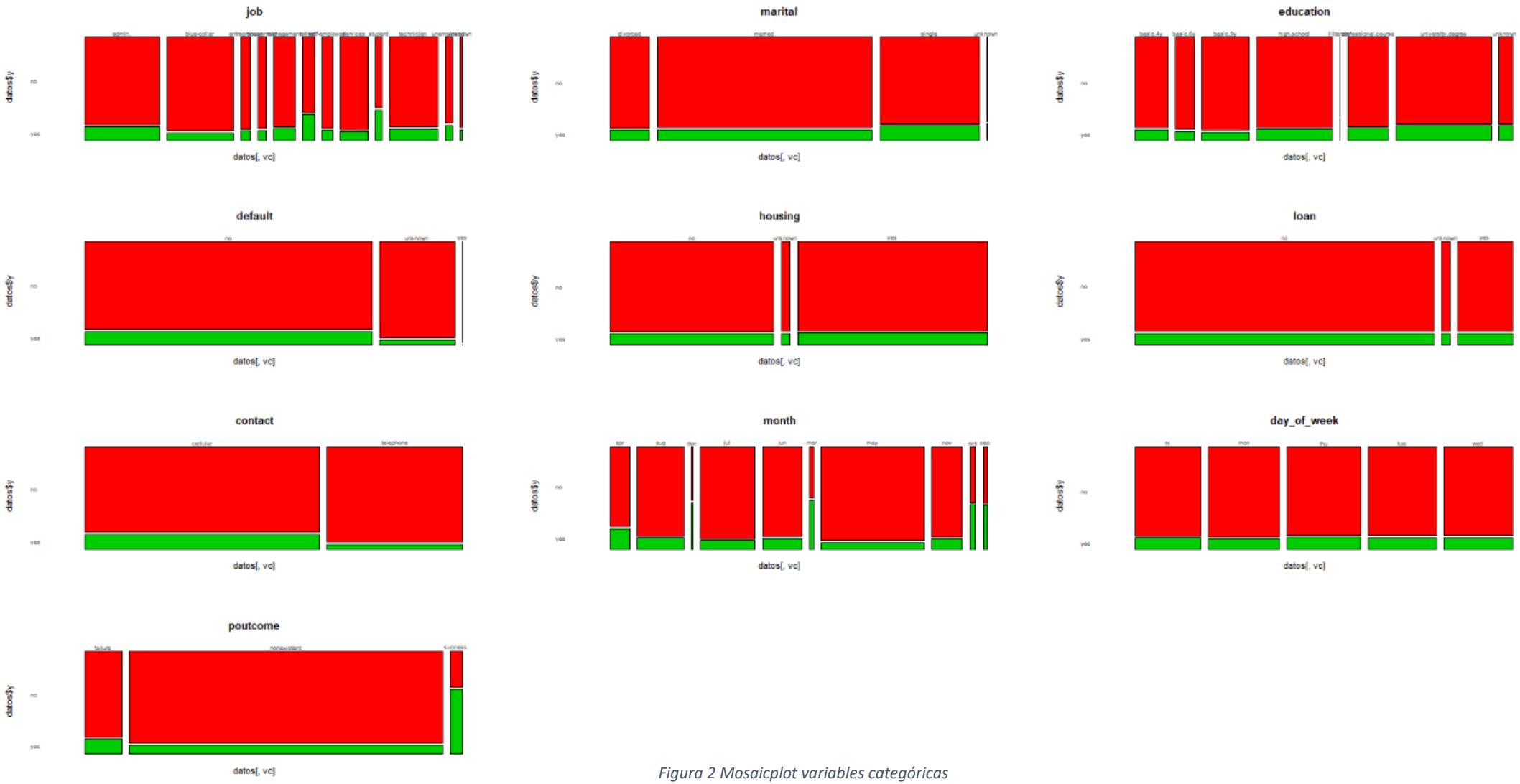



*Figura 2 Mosaicplot variables categóricas*

En el apartado 9 se pide juntar las categorías más similares, para ello primero hacemos un modelo solo incluyendo dicha variable.

```
Call:
glm(formula = datos$y ~ datos$education, family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6945  -0.5486  -0.4803  -0.4284   2.2586

Coefficients:
                                Estimate Std. Error z value   Pr(>|z|)
(Intercept)                     -2.13520    0.06075 -35.148    < 2e-16 ***
datos$educationbasic.6y         -0.20714    0.10656  -1.944   0.051912 .
datos$educationbasic.9y         -0.33430    0.08371  -3.994 0.00006510 ***
datos$educationhigh.school       0.03355    0.07232   0.464   0.642747
datos$educationilliterate        0.83592    0.65417   1.278   0.201307
datos$educationprofessional.course  0.12906    0.07963   1.621   0.105083
datos$educationuniversity.degree    0.31760    0.06839   4.644 0.00000342 ***
datos$educationunknown           0.33764    0.10231   3.300   0.000966 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20398  on 28644  degrees of freedom
Residual deviance: 20260  on 28637  degrees of freedom
AIC: 20276

Number of Fisher Scoring iterations: 5
```

*Figura 3 Modelo variable 'education'*

Una vez juntadas las categorías (ver script R) se consigue que dicha variable tenga un p-valor muy inferior.

```
Call:
glm(formula = datos$y ~ datos$education2, family = binomial,
    data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5351  -0.5351  -0.4615  -0.4615   2.1413

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.18617    0.02575 -84.891   <2e-16 ***
datos$education21  0.31475    0.03717   8.467   <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: 20398  on 28644  degrees of freedom
Residual deviance: 20327  on 28643  degrees of freedom
AIC: 20331

Number of Fisher Scoring iterations: 4
```

*Figura 4 Summary modelo glm00 con variable education2*

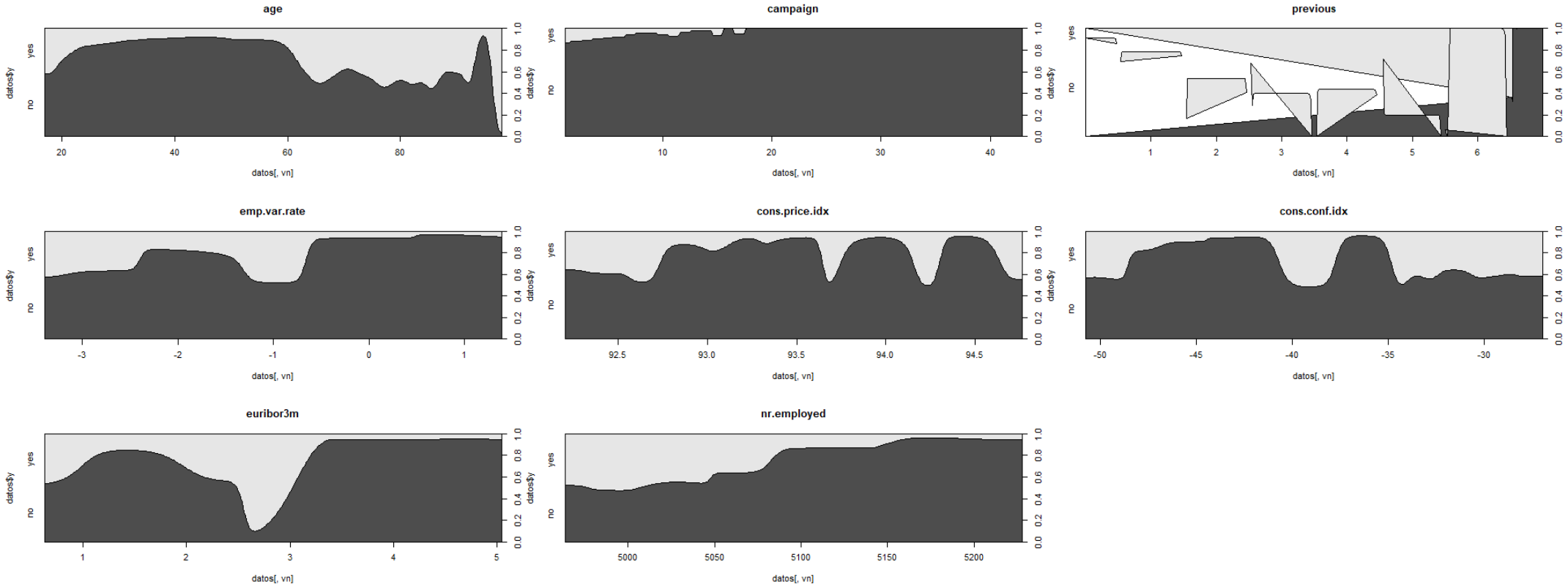



*Figura 5 Plot variables numéricas*

Construimos un modelo con todas las variables y nos da el siguiente resultado:

```
Call:
glm(formula = y ~ ., family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9608  -0.4123  -0.3261  -0.2684   2.9673

Coefficients:
                       Estimate  Std. Error z value Pr(>|z|)
(Intercept)        -179.510705   18.749592  -9.574  < 2e-16 ***
age                   0.002256    0.001909   1.182 0.237400
housingunknown       -0.060954    0.146801  -0.415 0.677983
housingyes           -0.034977    0.042435  -0.824 0.409800
contacttelephone     -0.934254    0.066453 -14.059  < 2e-16 ***
campaign             -0.049761    0.011106  -4.480 7.45e-06 ***
previous              0.075911    0.064747   1.172 0.241023
poutcomenonexistent   0.576798    0.103250   5.586 2.32e-08 ***
poutcomesuccess       1.740091    0.093179  18.675  < 2e-16 ***
emp.var.rate         -0.751719    0.071350 -10.536  < 2e-16 ***
cons.price.idx        1.526414    0.117404  13.001  < 2e-16 ***
cons.conf.idx         0.057205    0.006468   8.845  < 2e-16 ***
euribor3m            -0.332854    0.094614  -3.518 0.000435 ***
nr.employed           0.007452    0.001779   4.189 2.80e-05 ***
job21                -0.222135    0.063601  -3.493 0.000478 ***
month21              -0.659733    0.084891  -7.772 7.75e-15 ***
education21           0.176732    0.042887   4.121 3.77e-05 ***
default21            -0.349160    0.067679  -5.159 2.48e-07 ***
day_of_week21        -0.229383    0.052909  -4.335 1.46e-05 ***
marital21            -0.013589    0.044802  -0.303 0.761645
loan21                0.013814    0.058471   0.236 0.813233
```

En amarillo se destacan las variables no significativas para el modelo.

A continuación, se realiza una selección automática de las variables:

```
glm(formula = y ~ contact + campaign + poutcome + emp.var.rate +
    cons.price.idx + cons.conf.idx + euribor3m + nr.employed +
    job2 + month2 + education2 + default2 + day_of_week2, family = bin
    omial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9612  -0.4085  -0.3261  -0.2679   2.9752

Coefficients:
                       Estimate  Std. Error z value Pr(>|z|)
(Intercept)        -179.306088   18.739153  -9.569  < 2e-16 ***
contacttelephone     -0.941457    0.066197 -14.222  < 2e-16 ***
campaign             -0.049704    0.011107  -4.475 7.64e-06 ***
poutcomenonexistent   0.482427    0.064545   7.474 7.77e-14 ***
poutcomesuccess       1.754244    0.092359  18.994  < 2e-16 ***
emp.var.rate         -0.750719    0.071289 -10.531  < 2e-16 ***
cons.price.idx        1.534831    0.117279  13.087  < 2e-16 ***
cons.conf.idx         0.058048    0.006449   9.001  < 2e-16 ***
euribor3m            -0.329222    0.094550  -3.482 0.000498 ***
nr.employed           0.007300    0.001773   4.117 3.83e-05 ***
job21                -0.242948    0.061201  -3.970 7.20e-05 ***
month21              -0.658098    0.084785  -7.762 8.36e-15 ***
```

```
education21          0.173753   0.042715   4.068 4.75e-05 ***
default21           -0.341678   0.067067  -5.095 3.50e-07 ***
day_of_week21       -0.228468   0.052883  -4.320 1.56e-05 ***
```

En este modelo ya se observa que todas las variables incluidas son significativas. Hay algunas por eso con un p-valor claramente mayor al resto.

### 2.3 Hosmer and Lemeshow test + Validación gráfica

Este test es muy utilizado en regresión logísitca. Da información sobre la bondad de ajuste al m odelo que se propone. Dicho en otras palabras, si el modelo que se propone es capaz de explicar los que se observa. En este test evaluamos la distancia entre los valores observados y esperados.

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod.glm2$y, fitted(mod.glm2)
X-squared = 41.216, df = 8, p-value = 0.0000019
```

el p-valor es inferior a 0.05, por consiguiente, deberíamos descartar el modelo. El test de Hosmer y Lemeshow no siempre es el mejor indicativo de si un modelo se ajusta bien. Para la validación delo modelo también contamos con herramientas gráficas.

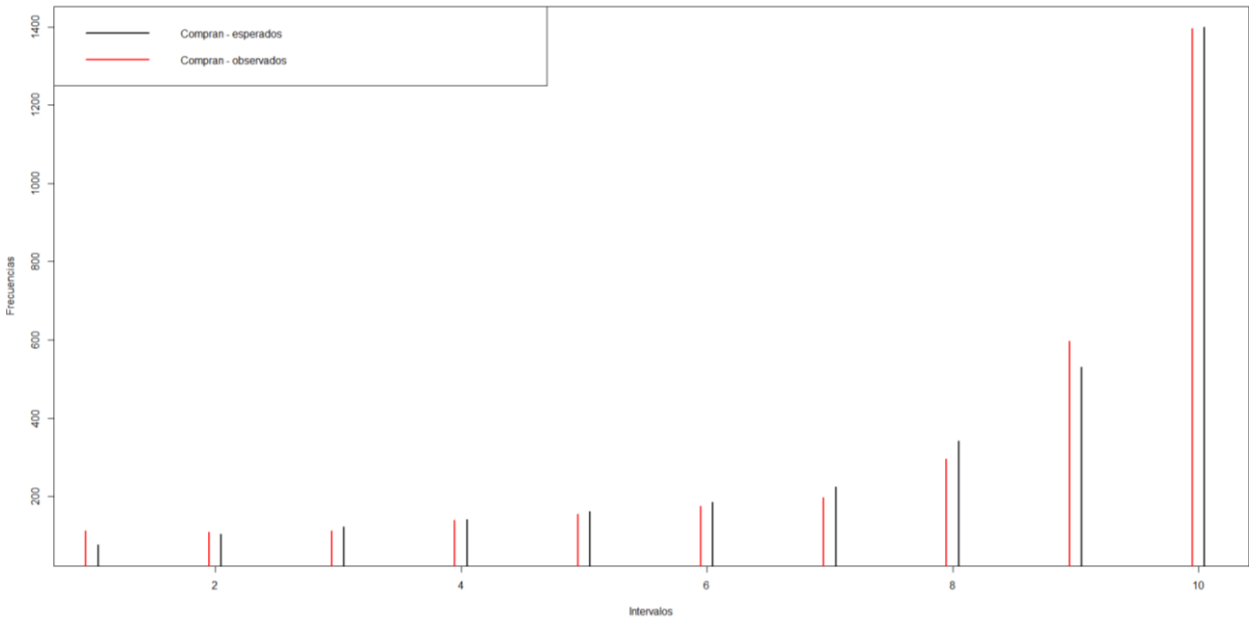



*Figura 6 Validación gráfica modelo*

Se observa en la figura 6, el gráfico de observados vs esperados. Con dicho gráfico queda patente que el modelo si se ajusta bien. La diferencia que se observa entre los valores observados y esperados a nivel gráfico no varían mucho unos de los otros.

```
        0%        10%        20%        30%        40%        50%        60%        70%        80%        90%       100%
0.00359518 0.03165941 0.04020120 0.04556437 0.05125821 0.05906479 0.07063577 0.09072235 0.14604721 0.27566842 0.86023595
```

*Figura 7 División de los intervalos según probabilidades*

Se realiza una división de las probabilidades según puntos de corte.

La formulación del modelo queda tal que:

*y ~ contact + campaign + poutcome + emp.var.rate + cons.price.idx + cons. conf.idx + euribor3m + nr.employed + job2 + month2 + education2 + default2 + day_of_week2*

No he realizado ninguna transformación sobre las variables numéricas. Las variables categóricas han sufrido las transformaciones mostradas en el script de R.

### 2.4 Odds ratio

El odd ratio de contratar vía teléfono fijo es de 0.39 mientras que hacerlo vía celular es de 2.56.

Eso quiere decir que la odds de venta de producto vía celular se incrementan por 6.56 veces respecto a la venta vía teléfono fijo.

$$OR = \frac{Odd_{celular}}{Odd_{fijo}}$$

### 2.5 Variable más relevante

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.00 | 0.00 |
| contacttelephone | 0.34 | 0.44 |
| campaign | 0.93 | 0.97 |
| poutcomenonexistent | 1.43 | 1.84 |
| poutcomesuccess | 4.83 | 6.93 |
| emp.var.rate | 0.41 | 0.54 |
| cons.price.idx | 3.68 | 5.84 |
| cons.conf.idx | 1.05 | 1.07 |
| euribor3m | 0.60 | 0.87 |
| nr.employed | 1.00 | 1.01 |
| job21 | 0.70 | 0.88 |
| month21 | 0.44 | 0.61 |
| education21 | 1.09 | 1.29 |
| default21 | 0.62 | 0.81 |
| day_of_week21 | 0.72 | 0.88 |

Observamos que la variable más relevante a la hora de tener éxito en la venta és la variable "poutcomesuccess".

## 3. CURVA ROC Y AUC

Hemos obtenido unos resultados en mi opinión muy correctos. Se aprecia que el modelo predice con una exactitud alta AUC = 0.748.

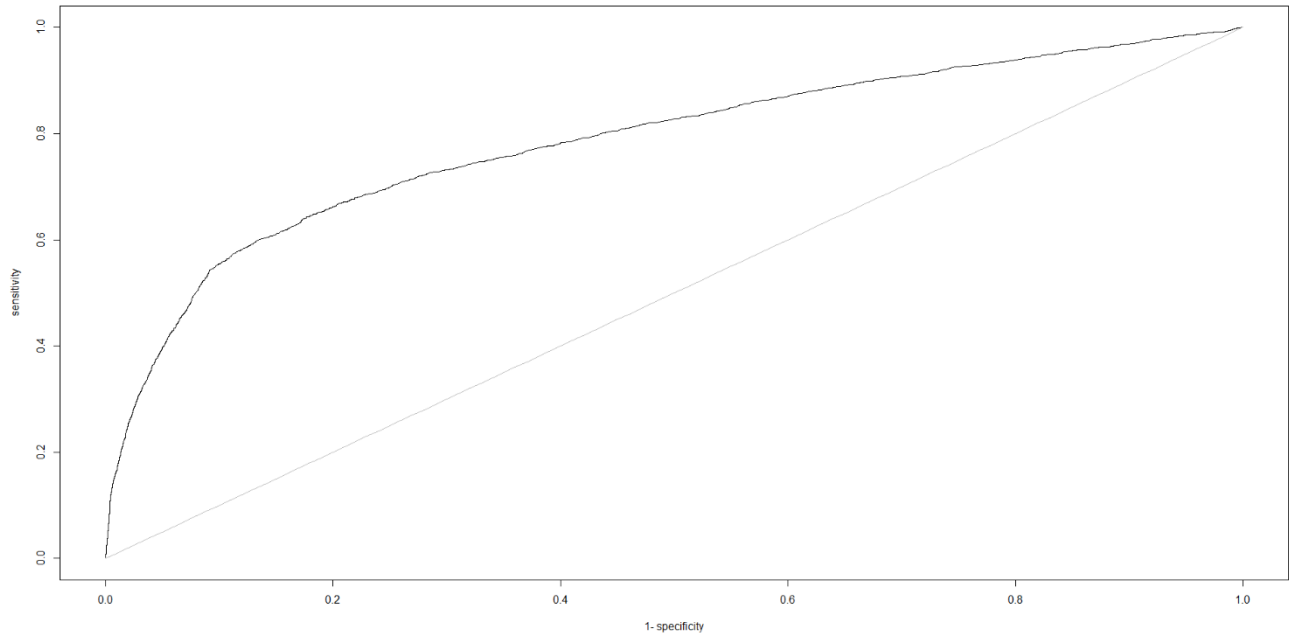



*Figura 8 Curva roc*

Como siempre, hay cierta complejidad en la interpretación del papel de las variables a la hora de predecir la respuesta.

En la matriz de confusión se observan los porcentajes de los resultados esperados cuando fijamos la probabilidad de adquirir el producto > 0.2.

| llamar | no | yes | llamar | no | yes |
|---|---|---|---|---|---|
| no | 23308 | 1628 | no | 93.5 | 6.5 |
| si | 2054 | 1655 | si | 55.4 | 44.6 |

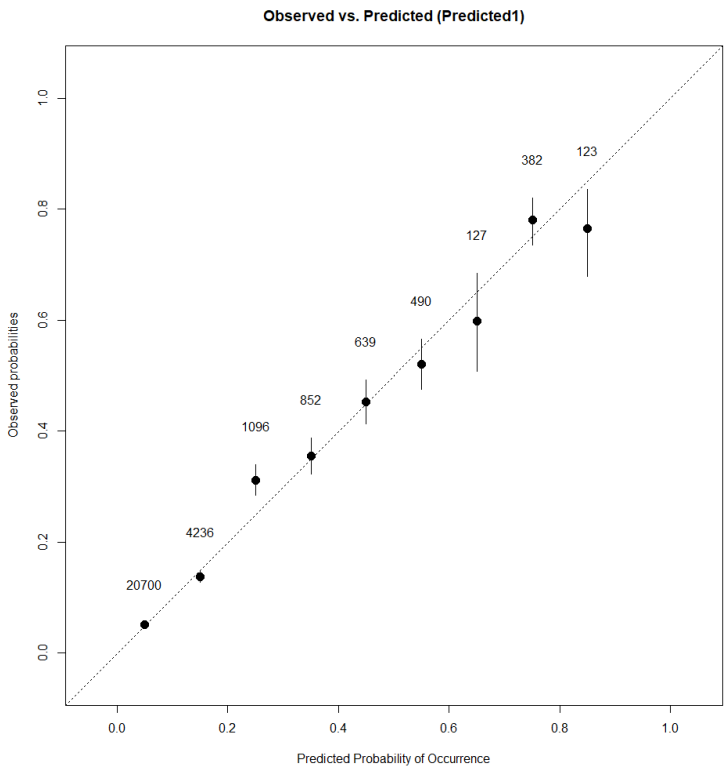



*Figura 9 Gráfico adicional: observed vs predicted*

Se observa claramente que gracias a las transformaciones realizadas el modelo ha mejorado de forma significativa.

Se ha realizado un profundo análisis del temario correspondiente al modelo lineal, lo que me ha ayudado a entender mejor los conceptos aprendidos con anterioridad en clase.