UNIVERSITAT OBERTA DE CATALUNYA
TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS
Autor: Jhon Gómez Higuera

1. Descripción del dataset. Por qué es importante y qué pregunta/problema pretende responder?

Los datos correponden a aquellos utilizados originalmente por [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001 y han sido tomados de Kaggle, los cuales a su vez tienen como fuente el UCI Machine Learning Repository. En consecuencia, los datos aquí utilizados pueden ser encontrados en: https://www.kaggle.com/henriqueyamahata/bank-marketing

Originalmente, el conjunto de datos está compuesto por 41.188 observaciones y 21 variables, una de las cuales corresponde a la variable objetivo y que indica si el individuo al que se realizó la campaña de marketing aceptó el producto, esto es, si aperturó o no la cuenta en el banco. La Tabla 1 presenta la información de las variables.

Nombre	Descripción	Tipo
age	Edad del individuo	Numérica
job	Tipo de empleo. Niveles: “admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown"	Categórica
marital	Estado civil. Niveles: “divorced","married","single","unknown"	Categórica
education	Nivel educativo. Niveles: “basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown"	Categórica
default	El individuo tiene créditos castigados? Niveles: Yes, No	Categórica
housing	El individuo tiene créditos de vivienda? Niveles: Yes, No	Categórica
loan	El individuo tiene créditos de consumo? Niveles: Yes, No	Categórica
contact	Canal del último contacto. Niveles: “cellular","telephone"	Categórica
month	Último mes de contacto	Categórica
day_of_week	Día de la semana del último contacto	Categórica
durtation	Duración, en segundos, del último contacto	Numérica
campaign	Número de veces que se contactó al cliente durante el desarrollo de la campaña de marketing	Numérica
pdays	Número de días desde la última campaña en que se contactó al cliente	Numérica
previous	Número de veces que se contactó al cliente antes de la actual campaña de marketing	Numérica
poutcome	Resultado de la anterior campaña de marketing. Niveles: “failure","nonexistent","success"	Categórica
emp.var.rate	Tasa trimestral del empleo	Numérica
cons.price.idx	Indice de precios del consumidor, mensual	Numérica
cons.conf.idx	Indice de confianza del consumidor, mensual	Numérica
euribor3m	Tasa Euribor a 3 meses, indicador diario	Numérica
nr.employed	Número de ocupados, trimestral	Numérica
y	El cliente ha aperturado el producto? Niveles: Yes, No	Numérica

Tabla 1. Detalle del conjunto de datos

Este conjunto de datos permite evaluar la efectividad de una campaña de marketing realizada por un banco a un conjunto de clientes, en términos de si los mismos adquirieron el producto ofertado en esta: un depósito a término, el cual se activaba de manera telefónica, motivo por el cual la variable “duration” no será considerada, pues en la misma se observaba el resultado qs positivo o no del cliente.

2. Análisis exploratorio

Tal como se indicó en la sección anterior, debido a que la duración del último contacto se desconoce, pero además en el mismo se sabe la respuesta del cliente, para efectos de un ejercicio riguroso, esta variable no es considerada en el análisis. Por demás, evidentemente una duración igual a cero tuvo un resultado de y igual a no.

2.1 Análisis de casos ausentes

```
      age            campaign          pdays          previous       emp.var.rate    
 Min.   :17.00   Min.   : 1.000   Min.   :  0.0   Min.   :0.000   Min.   :-3.40000  
 1st Qu.:32.00   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   1st Qu.:-1.80000  
 Median :38.00   Median : 2.000   Median :999.0   Median :0.000   Median : 1.10000  
 Mean   :40.02   Mean   : 2.568   Mean   :962.5   Mean   :0.173   Mean   : 0.08189  
 3rd Qu.:47.00   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000   3rd Qu.: 1.40000  
 Max.   :98.00   Max.   :56.000   Max.   :999.0   Max.   :7.000   Max.   : 1.40000  
```

```
cons.price.idx  cons.conf.idx     euribor3m      nr.employed   
 Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964  
 1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099  
 Median :93.75   Median :-41.8   Median :4.857   Median :5191  
 Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5167  
 3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228  
 Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228  
```

Imagen 1. Resumen de variables numéricas

Para el caso de las variables numéricas, la Imagen 1 resume los estadísticos básicos de distribución de las mismas. En particular, la variable pdays mantiene un alto registro de valores 999, caso para el cual esto es considerado como un dato ausente. Debido a su alta proporción respecto del total de datos, esta variable se descarta en el análisis.

De otro lado, en las variables categóricas, tal como lo muestra la Tabla 1, existe el valor “unknown”, para el cual la Imagen 2 resume tal información. Las variables job, marital, education, default, housing y loan registran valores ausentes donde el número de datos ausentes para la variable default es muy alto: 20,8% del total de las observaciones. La alta presencia de datos ausentes en esta variable es un elemento complejo, pues indicaría que los sistemas del banco no tendrían información precisa sobre el estado de los productos de sus clientes o que la misma fue capturada de fuentes no oficiales, entre otros motivos. Cualquiera que el sea, puesto que es difícil inferir si un cliente

contaba o no con un producto de crédito y además si en su historia registró castigos o incumplimiento total de los pagos, se eliminan todas aquellas observaciones con valor desconocido en esta variable.

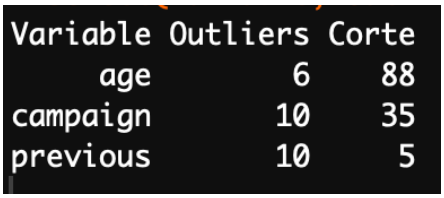
Imagen 2. Resumen de variables categóricas con NA

Lo anterior también se aplica para el caso de las variables loan, housing y education, casos en los cuales no se considera conveniente imputar valores a los datos ausentes. Para las variables job y marital se propone un tratamiento distinto a los datos ausentes. En el caso de la primera se asignarán los casos a la categoría desempleados, mientras que en el caso de marital serán asignados a la de mayor frecuencia: casados.

Así las cosas, con esta reducción de dimensiones, el conjunto de datos ahora tendría 30.667 observaciones y 19 atributos.

2.2 Análisis de datos atípicos (outliers)

El tratamiento de los datos atípicos parte de su identificación. Así, para el caso de las variables continuas se aplicará la función ronserTest(), incluida en el paquete EnvStats del lenguaje R de programación. La misma parte del supuesto de que la variable sigue una distribución Normal y encuentra los valores extremos. La Imagen 3 a continuación presenta el resultado del test para las variables age, campaign, previous, emp.var.rate, cons.price.index, cons.conf.idx, euribor3m y nr.employed.



Variable	Outliers	Corte
age	6	88
campaign	10	35
previous	10	5

Imagen 3. Resumen de datos atípicos para variables numéricas

Como lo muestra la imagen anterior, 3 variables registran valores extremos, para los cuales, además, se ha incluido el valor mínimo a partir del cual se eliminarán las observaciones. Esto último se hace debido a que el número de observaciones atípicas, comparado con el total de la muestra, es escaso.

2.3 Análisis de datos "limpios"

Como resultado de los procesos de limpieza de los datos realizado en las secciones anteriores, se cuenta con un conjunto de datos con 30.547 observaciones y 19 atributos. La Imagen 4 presenta el resumen de todas las variables.

```
     age                job           marital                   education     default     housing
Min.   :17.00   admin.     :8727   divorced: 3525   basic.4y          : 2383   no :30544   no :14015
1st Qu.:31.00   blue-collar:5677   married :17573   basic.6y          : 1403   yes:    3   yes:16532
Median :37.00   technician :5460   single  : 9449   basic.9y          : 4298
Mean   :38.99   services   :2857                    high.school       : 7696
3rd Qu.:45.00   management :2307                    illiterate        :   11
Max.   :86.00   retired    :1183                    professional.course: 4317
                (Other)    :4336                    university.degree  :10439
 loan            contact          month      day_of_week    campaign         previous            poutcome
no :25762   cellular :20457   may    :9778   fri:5737    Min.   : 1.000   Min.   :0.0000   failure    : 3451
yes: 4785   telephone:10090   jul    :5096   mon:6292    1st Qu.: 1.000   1st Qu.:0.0000   nonexistent:25936
                              aug    :4677   thu:6419    Median : 2.000   Median :0.0000   success    : 1160
                              jun    :3628   tue:5977    Mean   : 2.507   Mean   :0.1845
                              nov    :3491   wed:6122    3rd Qu.: 3.000   3rd Qu.:0.0000
                              apr    :2124               Max.   :33.000   Max.   :3.0000
                              (Other):1753
 emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m      nr.employed     y
Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634   Min.   :4964   no :26716
1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.313   1st Qu.:5099   yes: 3831
Median : 1.10000   Median :93.44   Median :-41.80   Median :4.856   Median :5191
Mean   :-0.06833   Mean   :93.52   Mean   :-40.59   Mean   :3.466   Mean   :5161
3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228
Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045   Max.   :5228
```

Imagen 4. Resumen del conjunto de datos pre-procesado

En particular, de la imagen anterior puede establecerse el bajo número de observaciones con default, apenas 3, así como que no existe un resultado para la campaña anterior (poutcome) con alta proporción, hecho que también puede evidenciarse en la variable previos. Puesto que el proceso de limpieza de datos es uno iterativo, estas dos variables no serán consideradas en el análisis posterior.

Los gráficos 1 y 2 permiten conocer la distribución de las variables. El Gráfico 1 muestra la distribución de las variables numéricas, en el que llama la atención la alta proporción del valor cero en las variables campaign y previous. De otro lado, del

Gráfico 2 puede evidenciarse que la mayor parte de individuos son casados, cuentan con grados universitarios, fueron contactados vía celular, primordialmente en el mes de mayo y sin variaciones aparentemente significativas en los días de la semana, con contactos decrecientes desde el lunes al viernes.

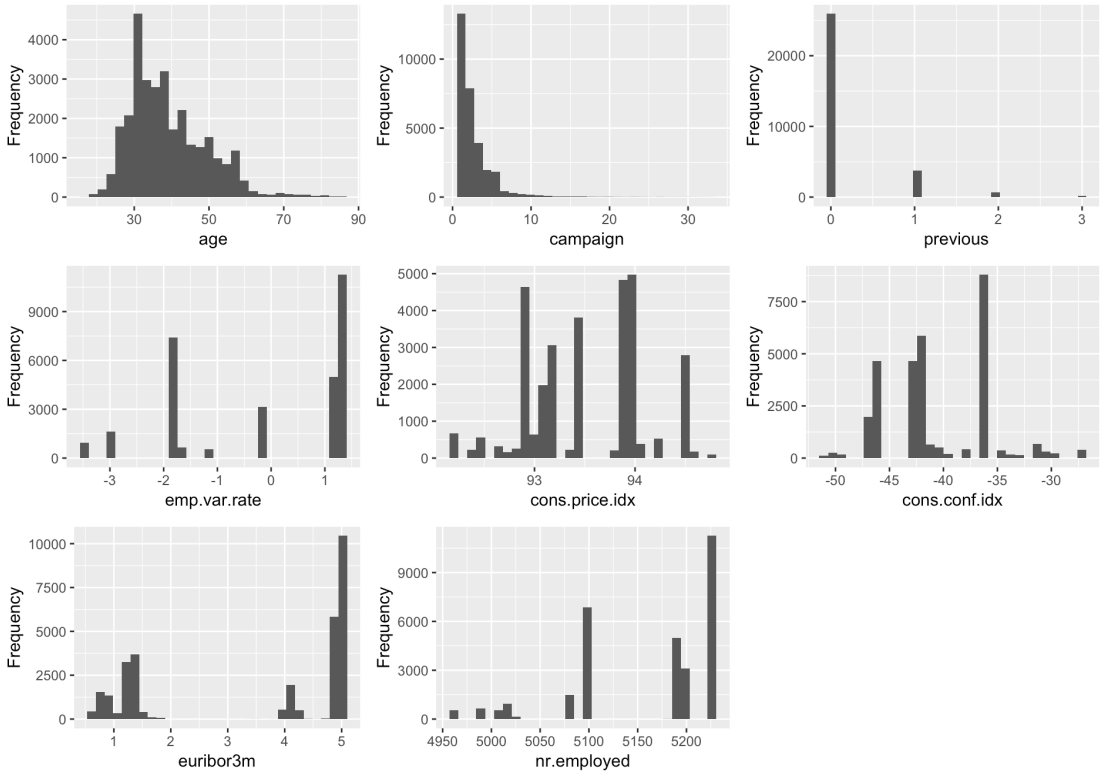


Gráfico 1. Distribución de las variables numéricas

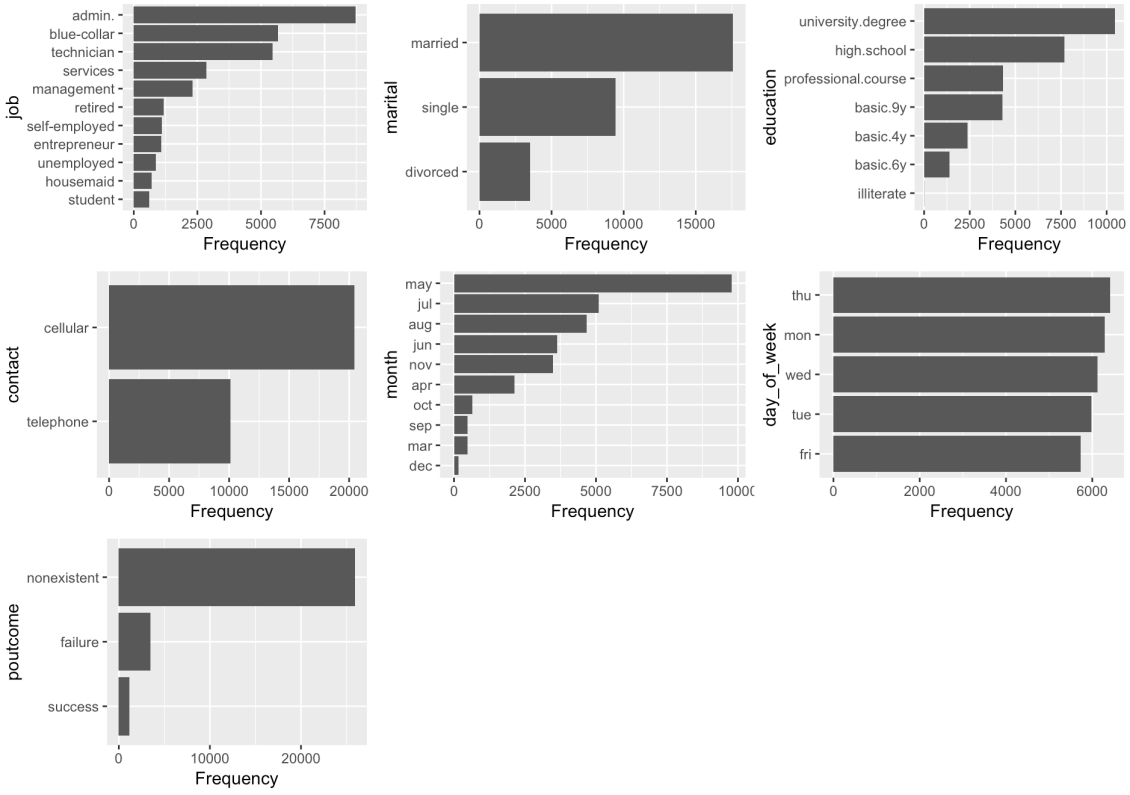


Gráfico 2. Distribución de las variables categóricas

Para efectos de seleccionar un conjunto de variables, a continuación también se presentan los resultados del análisis de correlaciones entre la variable objetivo y los atributos disponibles en el conjunto de datos. Es preciso considerar que este análisis se divide en función del tipo de variable y por tanto las imágenes a continuación presentan la correlación con las variables numéricas y con aquellas categóricas.

Variable	Coeficiente	p_Value
age	0.04	4.8104e-14
campaign	-0.07	0.0000e+00
previous	0.22	0.0000e+00
emp.var.rate	-0.30	0.0000e+00
cons.price.idx	-0.13	0.0000e+00
cons.conf.idx	0.06	0.0000e+00
euribor3m	-0.31	0.0000e+00
nr.employed	-0.36	0.0000e+00

Imagen 5. Correlación con las variables numéricas

La Imagen 5 presenta la correlación entre la variable Y y las variables numéricas que para el caso se ha obtenido mediante la función cor.test(), la cual permite a la vez verificar si el grado de asociación es estadísticamente significativo y por tanto se presenta información de la variable, el coeficiente y el valor de probabilidad del estadístico T-Student, cuya hipótesis es que la correlación no es estadísticamente

significativa. Puesto que los valores de probabilidad (p_Value) son menores a cualquier nivel de significancia alfa (0,05 para el caso), se rechaza la hipótesis mencionada y en consecuencia puede concluirse que la asociación entre cada atributo y la variable y tiene efecto estadístico significativo. No obstante, los valores del coeficiente son muy bajos, en algunos casos cercanos a cero, lo que podría sugerir una baja capacidad predictiva de los atributos respecto del resultado de la campaña de marketing.

La Imagen 6 resume la información del test Chi-Cuadrado de independencia entre las variables categóricas y la variable Y. Este test tiene como hipótesis el que las variables son independientes o que no existe correlación entre las mismas, donde la regla de contraste también consiste en comparar el valor del nivel de significancia alfa con el valor de probabilidad del estadístico. De esta imagen se concluye que no existe correlación con default, con loan, ni con housing.

Variable	p_Value
job	0.000000e+00
marital	6.972000e-14
education	0.000000e+00
default	1.000000e+00
housing	9.492832e-02
loan	3.274461e-01
contact	0.000000e+00
month	0.000000e+00
day_of_week	9.553606e-05
poutcome	0.000000e+00

Imagen 6. Test Chi-cuadrado de correlación para variables

Finalmente, el Gráfico 3 presenta información de la densidad de las variables numéricas según cada caso en la variable objetivo. En particular, puede observarse que no hay diferencias marcadas en la variable campaign, mientras que en las demás variables se puede observar distribuciones distintas.

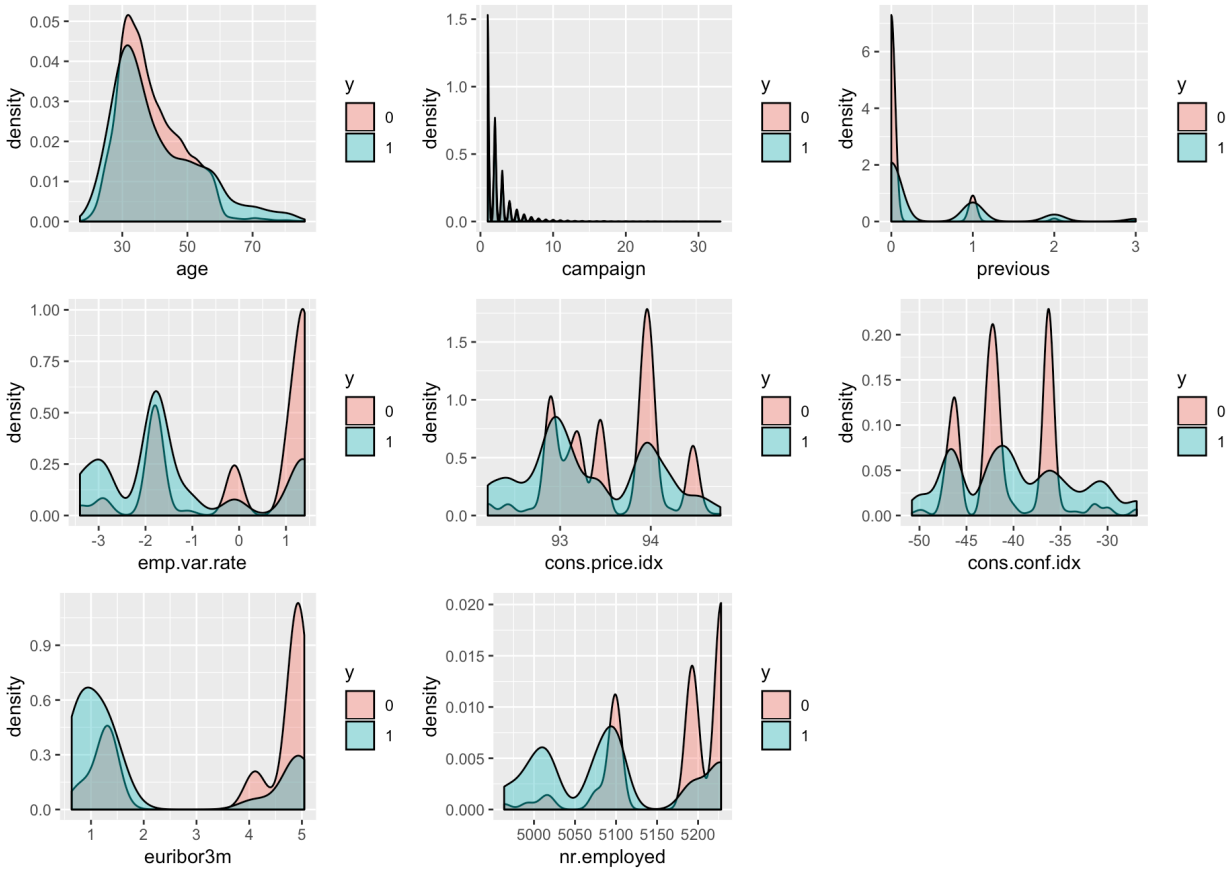


Gráfico 3. Distribución de las variables numéricas de acuerdo a la variable objetivo

2.4 Pruebas de normalidad y transformación de los datos

En las secciones anteriores se ha hecho la limpieza y reducción de dimensiones de los datos, no obstante, puesto que las unidades de medida de las variables numéricas es distinta y esto puede afectar los resultados al momento de aplicación de un algoritmo, a continuación se presentan los análisis de las variables normalizadas y su relación con la variable objetivo.

Para establecer si las variables numéricas siguen una distribución Normal se aplicaría un test como el de Shapiro Wilks. Sin embargo, debido al tamaño de la muestra con que se cuenta, se puede asimilar que las variables numéricas siguen una distribución Normal. Aún así, el Gráfico 4 presenta el gráfico de cuantiles para las variables: "age", "campaign", "previous", "emp.var.rate", "cons.price.idx", "cons.conf.idx", "euribor3m", “nr.employed”, respectivamente de izquierda a derecha, iniciando en la parte superior.

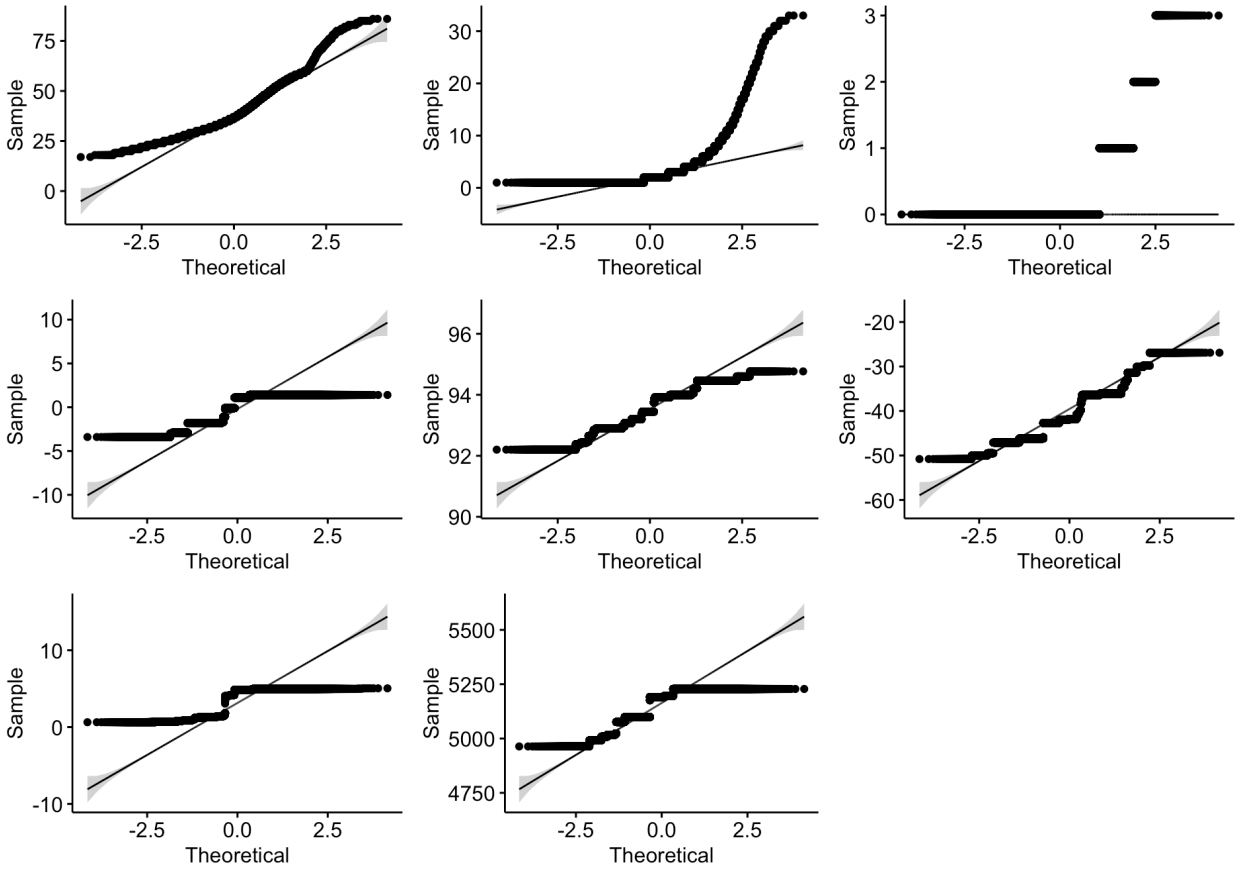


Gráfico 4. Gráfico de cuantiles para análisis de normalidad

3. Modelado

Uno de los elementos a considerar en el problema del éxito de la campaña de marketing es el hecho de contar con clases no balanceadas. El Gráfico 5 permite conocer esta situación.

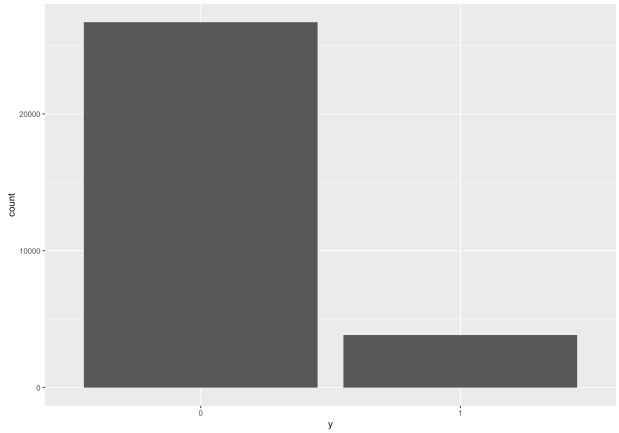


Gráfico 5. Distribución de la variable objetivo

De forma preliminar, se ha estimado un modelo de regresión logístico cuyos resultados se presentan en la Imagen 7. Para el caso, se han incluido las variables “age", “job”, “marital”, “education”, “contact”, “month”, “day_of_week”, “campaign”, “previous”, “poutcome”, “emp.var.rate”, “cons.price.idx”, “cons.conf.idx”, “euribor3m" y “nr.employed”.

En general, se puede concluir que el tipo de ocupación que desempeña el individuo objeto de la campaña de marketing no determina qué posible es que acepte apertura el producto, excepto para el caso de los servicios y los pensionados. Lo mismo sucede con el estado civil, donde el mismo no tiene efectos significativos sobre la decisión de apertura o no el producto, al igual que el número de veces que había sido contacto el cliente de manera previa. La edad tampoco constituye un factor relevante para decidir si se aperturará o no el producto, hecho que se podría percibir desde el análisis de correlaciones, donde la misma era cercana a cero.

El canal de contacto, el mes, el día de la semana, el resultado de la anterior campaña y las variables económicas sí resultan factores determinantes de la decisión del cliente en tanto a aperturar el producto. Un modelo con estos factores se resume en la Imagen 8.

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.7719556  0.1996610 -13.883  < 2e-16 ***
age                          -0.0208001  0.0296268  -0.702 0.482635
`jobblue-collar`             -0.1494181  0.0928268  -1.610 0.107475
jobentrepreneur              -0.0738778  0.1400793  -0.527 0.597916
jobhousemaid                 -0.1004403  0.1800072  -0.558 0.576859
jobmanagement                 0.0085674  0.0966896   0.089 0.929394
jobretired                    0.3385274  0.1289513   2.625 0.008659 **
`jobself-employed`           -0.0398652  0.1317778  -0.303 0.762257
jobservices                  -0.2011441  0.1007647  -1.996 0.045915 *
jobstudent                    0.2335719  0.1370821   1.704 0.088403 .
jobtechnician                -0.0218472  0.0809083  -0.270 0.787141
jobunemployed                -0.1440980  0.1404798  -1.026 0.305007
maritalmarried               -0.0006286  0.0796818  -0.008 0.993705
maritalsingle                 0.0312222  0.0896219   0.348 0.727557
educationbasic.6y             0.2233837  0.1444611   1.546 0.122026
educationbasic.9y            -0.0183775  0.1156089  -0.159 0.873699
educationhigh.school          0.0672284  0.1115567   0.603 0.546749
educationilliterate           1.1921477  0.7606344   1.567 0.117043
educationprofessional.course  0.1083801  0.1205271   0.899 0.368537
educationuniversity.degree    0.1394331  0.1113014   1.253 0.210296
contacttelephone             -0.8152518  0.0878426  -9.281  < 2e-16 ***
monthaug                      0.4339630  0.1400325   3.099 0.001942 **
monthdec                      0.2136098  0.2460172   0.868 0.385246
monthjul                     -0.0133021  0.1106819  -0.120 0.904338
monthjun                     -0.8756112  0.1430983  -6.119 9.42e-10 ***
monthmar                      1.4978944  0.1707170   8.774  < 2e-16 ***
monthmay                     -0.3584712  0.0932695  -3.843 0.000121 ***
monthnov                     -0.5587634  0.1387531  -4.027 5.65e-05 ***
monthoct                      0.1085650  0.1774839   0.612 0.540743
monthsep                      0.3171073  0.2098584   1.511 0.130775
day_of_weekmon               -0.1313502  0.0782264  -1.679 0.093132 .
day_of_weekthu                0.1939513  0.0749766   2.587 0.009687 **
day_of_weektue                0.1646757  0.0774112   2.127 0.033396 *
day_of_weekwed                0.2542352  0.0769554   3.304 0.000954 ***
campaign                     -0.1114205  0.0328689  -3.390 0.000699 ***
previous                      0.0813533  0.0428076   1.900 0.057375 .
poutcomenonexistent           0.7364117  0.1312007   5.613 1.99e-08 ***
poutcomesuccess               1.7905298  0.1028292  17.413  < 2e-16 ***
emp.var.rate                 -2.8053057  0.2581039 -10.869  < 2e-16 ***
cons.price.idx                1.3586036  0.1706666   7.961 1.71e-15 ***
cons.conf.idx                 0.1219674  0.0441517   2.762 0.005737 **
euribor3m                     0.5636677  0.2762323   2.041 0.041295 *
nr.employed                   0.6192754  0.2740085   2.260 0.023818 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16149  on 21383  degrees of freedom
Residual deviance: 12632  on 21341  degrees of freedom
AIC: 12718
```

Imagen 7. Resultados modelo LOGIT

Puesto que el proceso de análisis de los datos es uno iterativo, también se ha realizado un análisis de componentes principales, pues del modelo en la Imagen 7 resulta un conjunto de datos con 10 atributos y una variable objetivo. No obstante, tal como lo enseña el Gráfico 6, el porcentaje de varianza explicado no es suficiente para considerar una reducción de dimensiones en el conjunto de datos.

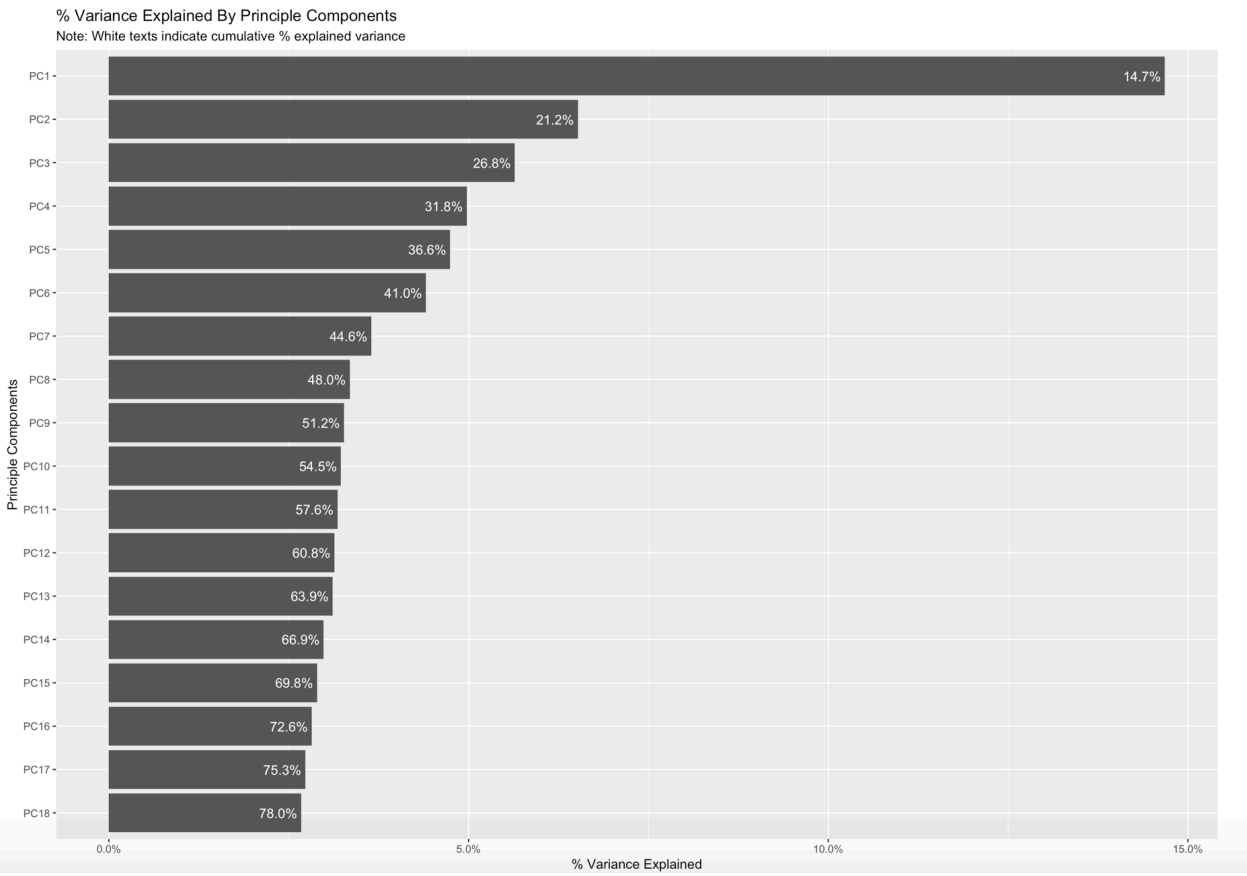


Gráfico 6. Análisis de componentes principales

De acuerdo con lo anterior, los resultados que se registran en la Imagen 8 considerar los 10 atributos seleccionados.

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.56223    0.12148 -21.093  < 2e-16 ***
contacttelephone   -0.78283    0.08770  -8.926  < 2e-16 ***
monthaug            0.65995    0.14027   4.705 2.54e-06 ***
monthdec            0.55429    0.24637   2.250 0.024459 *
monthjul            0.10878    0.11075   0.982 0.326002
monthjun           -0.74322    0.14334  -5.185 2.16e-07 ***
monthmar            1.70854    0.16940  10.086  < 2e-16 ***
monthmay           -0.36427    0.09356  -3.893 9.89e-05 ***
monthnov           -0.36748    0.13902  -2.643 0.008209 **
monthoct            0.24136    0.17736   1.361 0.173565
monthsep            0.24284    0.21129   1.149 0.250419
day_of_weekmon     -0.13950    0.07768  -1.796 0.072500 .
day_of_weekthu      0.19393    0.07451   2.603 0.009245 **
day_of_weektue      0.08419    0.07739   1.088 0.276619
day_of_weekwed      0.25584    0.07644   3.347 0.000817 ***
campaign           -0.06560    0.03172  -2.068 0.038641 *
poutcomenonexistent  0.46606    0.07064   6.598 4.17e-11 ***
poutcomesuccess     1.65855    0.10085  16.446  < 2e-16 ***
emp.var.rate       -2.71573    0.25900 -10.485  < 2e-16 ***
cons.price.idx      1.39365    0.16938   8.228  < 2e-16 ***
cons.conf.idx       0.13892    0.04342   3.200 0.001375 **
euribor3m           0.32437    0.27447   1.182 0.237284
nr.employed         0.68965    0.27117   2.543 0.010984 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16149  on 21383  degrees of freedom
Residual deviance: 12696  on 21361  degrees of freedom
AIC: 12742
```

Imagen 8. Resultados modelo LOGIT ajustado

4. Evaluación de resultados y diagnóstico

Una vez se ha propuesto un modelo aún quedan preguntas por responder: cuál es el ajuste del modelo a los datos?, cuál atributo es el más importante? y cuál es el grado de precisión?, entre otras cuestiones. Esta sección presenta algunos elementos que ayudan a aclarar estas preguntas.

4.1 Bondad de ajuste

Para el caso del modelo LOGIT se estimará la razón de verosimilitud, la cual busca establecer si el modelo con menos atributos es mejor a aquél que incluyó la totalidad de estos. La conclusión que se extrae de la Imagen 9 es que el modelo que incluye el total de variables o “sin restricciones” es mejor.

```
Likelihood ratio test

Model 1: y ~ age + job + marital + education + contact + month + day_of_week +
    campaign + previous + poutcome + emp.var.rate + cons.price.idx +
    cons.conf.idx + euribor3m + nr.employed
Model 2: y ~ contact + month + day_of_week + campaign + poutcome + emp.var.rate +
    cons.price.idx + cons.conf.idx + euribor3m + nr.employed
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  43 -6338.9
2  23 -6306.9 -20 63.879  1.758e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Imagen 9. Prueba de la razón de verosimilitud para los modelos

Una vez extraída la conclusión de la Imagen 9, la Imagen 10 presenta el seudo R2 cuyos valores oscilan entre 0 y 1, donde valores cercanos a cero indican que el modelo no tiene capacidad predictiva. Para el caso, el indicador de McFadden es 0.213

llh	llhNull	G2	McFadden	r2ML	r2CU
-6338.8782660	-8054.7735231	3431.7905142	0.2130284	0.1482750	0.2801723

Imagen 10. Seudo R2

La importancia relativa de los atributos se presenta en la Imagen 11, donde de nuevo es evidente que el canal de contacto, el mes de aplicación de la campaña, el resultado de la campaña previa a la actual y las variables económicas son los principales atributos que explican la decisión del cliente de aceptar la oferta del banco.

	Overall
age	1.14697892
jobblue-collar	1.90618605
jobentrepreneur	2.00717455
jobhousemaid	0.79849781
jobmanagement	0.04009634
jobretired	2.31034210
jobself-employed	0.75553861
jobservices	2.31035205
jobstudent	1.47852696
jobtechnician	0.14842024
jobunemployed	1.00769973
maritalmarried	0.38312803
maritalsingle	0.58872033
educationbasic.6y	0.04556591
educationbasic.9y	0.57245069
educationhigh.school	0.80312689
educationilliterate	1.58336735
educationprofessional.course	0.53681421
educationuniversity.degree	0.62589155
contacttelephone	9.02990660
monthaug	3.74055775
monthdec	2.37831562
monthjul	0.18047206
monthjun	5.80538578
monthmar	9.62940431
monthmay	4.92348264
monthnov	3.49712479
monthoct	0.07548988
monthsep	2.01749032
day_of_weekmon	2.47102253
day_of_weekthu	1.81812298
day_of_weektue	1.80759429
day_of_weekwed	2.91610516
campaign	2.76925688
previous	1.72902670
poutcomenonexistent	5.17223744
poutcomesuccess	16.14162150
emp.var.rate	10.70462652
cons.price.idx	8.50442578
cons.conf.idx	3.31691486
euribor3m	1.19518910
nr.employed	2.80597869

Imagen 11. Importancia relativa de los atributos

La Imagen 12 permite conocer la matriz de confusión del modelo seleccionado, aplicado a los datos de prueba, donde se establece que el modelo tiene una precisión de 89%.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7891  848
         1  151  274

               Accuracy : 0.891
                 95% CI : (0.8844, 0.8973)
    No Information Rate : 0.8776
    P-Value [Acc > NIR] : 3.684e-05

                  Kappa : 0.3077
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9812
            Specificity : 0.2442
         Pos Pred Value : 0.9030
         Neg Pred Value : 0.6447
             Prevalence : 0.8776
         Detection Rate : 0.8611
   Detection Prevalence : 0.9536
      Balanced Accuracy : 0.6127

       'Positive' Class : 0
```

Imagen 12. Matriz de confusión

5. Resolución del problema. A partir de los resultados obtenidos, cuáles son las conclusiones? Los resultados permiten responder al problema?

Como conclusión, puede considerarse que la efectividad de la campaña de marketing del banco depende primordialmente de los factores económicos del entorno del individuo, del mes y día de contacto y del canal por el cual se hace el mismo, así como del resultado de la campaña inmediatamente anterior. Esto cambia las posibilidades de que el individuo aperture el producto.

(Intercept)	age	`jobblue-collar`	jobentrepreneur
0.07648013	0.96671425	0.83646293	0.73869275
jobhousemaid	jobmanagement	jobretired	`jobself-employed`
0.86752617	1.00388370	1.34155314	0.90396503
jobservices	jobstudent	jobtechnician	jobunemployed
0.79356909	1.22109739	0.98811272	0.87122447
maritalmarried	maritalsingle	educationbasic.6y	educationbasic.9y
0.97061134	0.94941186	1.00698119	0.93601543
educationhigh.school	educationilliterate	educationprofessional.course	educationuniversity.degree
1.09304561	3.38175026	1.06678860	1.07209978
contacttelephone	monthaug	monthdec	monthjul
0.45843205	1.69344423	1.80278485	0.98034776
monthjun	monthmar	monthmay	monthnov
0.43163631	4.98247304	0.63122879	0.61554564
monthoct	monthsep	day_of_weekmon	day_of_weekthu
1.01349383	1.52121088	0.82557619	1.14529209
day_of_weektue	day_of_weekwed	campaign	previous
1.14749995	1.24818871	0.91736783	1.07578964
poutcomenonexistent	poutcomesuccess	emp.var.rate	cons.price.idx
1.95871518	5.23479297	0.06175728	4.24159836
cons.conf.idx	euribor3m	nr.employed	
1.15544494	1.38707164	2.14157236	

Imagen 13. Coeficientes del modelo

De acuerdo con la Imagen 13, hay mayores posibilidades de que un individuo acepte la oferta del banco si el contacto se da en marzo, vía teléfono y si, por ejemplo, el resultado de la campaña anterior fue exitoso con el mismo individuo.

6. Código y datos resultantes

El código en lenguaje R puede ser obtenido en la url:
https://github.com/jgomezhiguera/BankMarketing/blob/master/Code/Bank_Script.R

Del mismo modo, los datos resultantes se encuentran disponibles en la url:
https://github.com/jgomezhiguera/BankMarketing/tree/master/Clean%20data