MINERIA DE DATOS CON SAS

-

BREVE DESCRIPCIÓN DE LA PRÁCTICA:

MODELO DE VENTA CRUZADA EN BANCO

1. Introducción

El equipo de marketing de un banco quiere hacer una campaña para vender un depósito bancario.

Para predecir el éxito de la campaña de marketing telemático habrá que hacer un modelo que nos permita distinguir que clientes son los más propensos a la compra del producto.

Para abordar la practica voy a seguir las practicas de preprocesamiento que se emplearon en el siguiente trabajo académico: <http://support.sas.com/resources/papers/proceedings17/2029-2017.pdf>.

Sin embargo voy a modelizar de diferente manera y valorar modelos alternativos para atacar problemas de este tipo.

2. Datos

El conjunto de datos es el conjunto de datos Marketing del Banco Portugués en la Universidad de California, Irvine (UCI) máquina de aprendizaje repositorio situado en la siguiente URL:[https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.](https://www.microsofttranslator.com/bv.aspx?from=en&to=es&a=https%3A%2F%2Farchive.ics.uci.edu%2Fml%2Fdatasets%2FBank%2BMarketing)Los datos son el resultado de una campaña de marketing directo realizada por una entidad bancaria portuguesa para vender a plazo depósitos o certificado de depósitos. La institución bancaria realizaron llamadas telefónicas a potenciales compradores desde mayo de 2008 a noviembre de 2010. Se utilizó el conjunto de datos completo, bank\_additional\_full.csv.

Hay 41.188 observaciones 21 Variables y en el conjunto de datos. Hay 10 variables de medida continua y 10 variables categóricas. La respuesta del destino (y) es una respuesta binaria que indica si el cliente suscrito a un depósito a plazo o no. 'Sí' (valor numérico 1) indica el cliente suscrito a un depósito a plazo. 'No' (valor numérico 0) indica que el cliente no suscribían depósitos a plazo. Las variables se dividen en 4 categorías: datos de los clientes, ultima información del contacto, otros, variables sociales y económicas.

\*\*\* La variable ‘duration’ se muestra en la tabla 1 pero no se utilizará en el análisis debido al alto impacto sobre la respuesta del destino (y) por la descripción de la variable.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Variable |  |  | Categoría variable |  |  | Descripción |  |  | Variable |  |
|  |
|  |  |  |  |  |  |  |  |  |  | Tipo |  |
|  | age |  |  | Datos del cliente |  |  | Edad de clientes a la hora de llamada |  |  | Continua |  |
|  | job | |  | Datos del cliente |  |  | Tipo de clientes de trabajo - admin '.',' blue - | |  | Categóricos | |
|  |  |  |  |  |  |  | collar', 'emprendedor', 'criada', 'gestión', 'jubilado',' auto - | |  |  |  |
|  |  |  |  |  |  |  | empleados ',' servicios 'estudiante', 'técnico', 'parado', 'desconocido') | |  |  |  |
|  | marital |  |  | Datos del cliente |  |  | Estado civil de clientes en el momento de la llamada - 'divorciado', 'casado', 'solo', 'desconocido'; Nota: |  |  | Categóricos |  |
|  |  |  |  |  |  |  | significa 'divorciada' divorciado o viudo |  |  |  |  |
|  | Education | |  | Datos del cliente |  |  | Formación de clientes en el momento de la llamada- | |  | Categóricos | |
|  |  |  |  |  |  |  | ' basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','universit | |  |  |  |
|  |  |  |  |  |  |  | y.degree','unknown' | |  |  |  |
|  | default |  |  | Datos del cliente |  |  | ¿Cliente tiene crédito en mora? -'no', 'yes', 'desconocido' |  |  | Categóricos |  |
|  | housing | |  | Datos del cliente |  |  | ¿El cliente tiene un préstamo de casa? -'no', 'yes', 'desconocido' | |  | Categóricos | |
|  | loan |  |  | Datos del cliente |  |  | ¿El cliente tiene un préstamo personal? -'no', 'yes', 'desconocido' |  |  | Categóricos |  |
|  | contact | |  | Últimos datos |  |  | Tipo de comunicación con el cliente – 'celular', 'teléfono' | |  | Categóricos | |
|  | month |  |  | Últimos datos |  |  | El mes pasado contacto del año con el cliente - 'jan', 'febrero', 'mar',..., 'noviembre', 'dec' |  |  | Categóricos |  |
|  | day\_of\_week | |  | Últimos datos |  |  | Último día de contacto de semana con el cliente - 'LUN', 'mar', 'Mie', 'Jue', 'Vie' | |  | Categóricos | |
|  | duration |  |  | Últimos datos |  |  | Última Contacta con duración, en segundos al cliente. Nota importante: este atributo altamente |  |  | Continua |  |
|  |  |  |  |  |  |  | afecta el destino de los resultados (por ejemplo, si duración = 0 entonces y = 'no'). Sin embargo, la duración no es |  |  |  |  |
|  |  |  |  |  |  |  | conocido antes de una llamada se realiza. También, después del final de la llamada y es obviamente |  |  |  |  |
|  |  |  |  |  |  |  | conocido. Así, esta entrada sólo debe ser incluida para fines de referencia y |  |  |  |  |
|  |  |  |  |  |  |  | debe ser descartado si la intención es tener un modelo predictivo realista. |  |  |  |  |
|  | campaign | |  | Otros |  |  | Número de contactos realizados durante esta campaña para este cliente (incluye ultima | |  | Continua | |
|  |  |  |  |  |  |  | en contacto con) | |  |  |  |
|  | pdays |  |  | Otros |  |  | Número de días que pasa después de pasado el cliente contactado de una anterior |  |  | Continua |  |
|  |  |  |  |  |  |  | campaña (numérico; 999 significa cliente no contactó previamente) |  |  |  |  |
|  | previous | |  | Otros |  |  | Número de contactos realizados antes de esta campaña y para el cliente | |  | Continua | |
|  | poutcome |  |  | Otros |  |  | Resultado de la campaña de comercialización anterior - 'fracaso', 'inexistente', 'éxito' |  |  | Categóricos |  |
|  | emp\_var\_rate | |  | Social y económico |  |  | Tasa de variación de empleo - indicador trimestral | |  | Continua | |
|  | cons\_price\_idx |  |  | Social y económico |  |  | Índice de precios al consumidor – indicador mensual; Índice de precios mensual al consumidor o IPC |  |  | Continua |  |
|  |  |  |  |  |  |  | mide los cambios en los precios pagados por los consumidores por una canasta de bienes y |  |  |  |  |
|  |  |  |  |  |  |  | servicios de cada mes. |  |  |  |  |
|  | cons\_conf\_idx | |  | Social y económico |  |  | Índice de confianza del consumidor – indicador mensual; En Portugal, el consumidor | |  | Continua | |
|  |  |  |  |  |  |  | el índice de confianza se basa en entrevistas con los consumidores sobre sus percepciones de | |  |  |  |
|  |  |  |  |  |  |  | la situación del país actual y el futuro económico y sus tendencias a | |  |  |  |
|  |  |  |  |  |  |  | compra. Se calcula mediante la diferencia entre la proporción de positivos | |  |  |  |
|  |  |  |  |  |  |  | las respuestas de evaluación y evaluación negativa las respuestas, pero no incluyen la | |  |  |  |
|  |  |  |  |  |  |  | proporción de respuestas neutrales. | |  |  |  |
|  | nr\_employed |  |  | Social y económico |  |  | Euribor 3 meses tasa – Indicador diario; Euribor es corto para Euro Interbank |  |  | Continua |  |
|  |  |  |  |  |  |  | Tarifa. Las tarifas de Euribor se basan en las tasas de interés promedio en que una gran |  |  |  |  |
|  |  |  |  |  |  |  | panel de los bancos europeos pedir prestado fondos de uno a otro que maduran después de 3 |  |  |  |  |
|  |  |  |  |  |  |  | meses. |  |  |  |  |
|  | Nr.employed | |  | Social y económico |  |  | Número de empleados – indicador trimestral; Número de personas empleadas para un | |  | Continua | |
|  |  |  |  |  |  |  | cuarto. | |  |  |  |
|  | y |  |  | Destino/respuesta |  |  | ¿El cliente ha suscrito un depósito a plazo? -'sí', 'no' |  |  | Categóricos / |  |
|  |  |  |  |  |  |  |  |  |  | Binario |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

3. Problema/objetivo

El primer objetivo de este estudio es determinar que variables tienen la mayor influencia sobre si un cliente compra un depósito a plazo o no. El segundo objetivo es determinar los niveles de las variables que producen el depósito a plazo más compras.

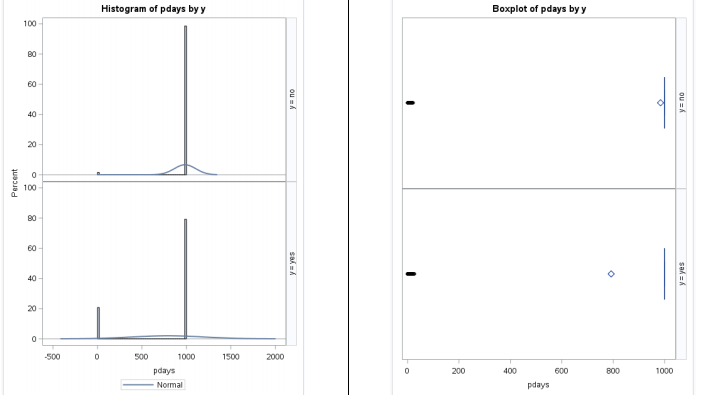
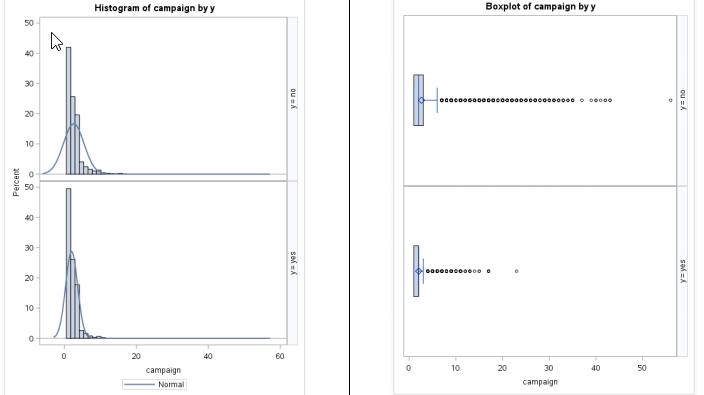
4. L impieza y validación de datos

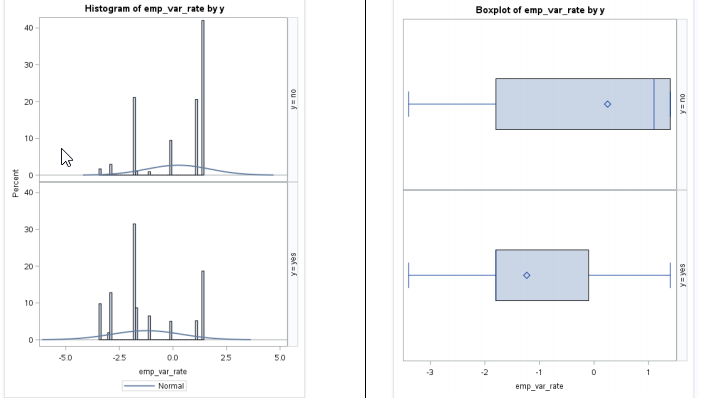
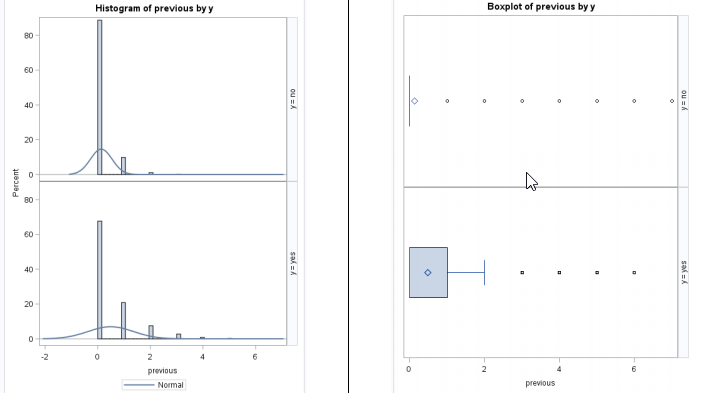
4.1. Variables continuas

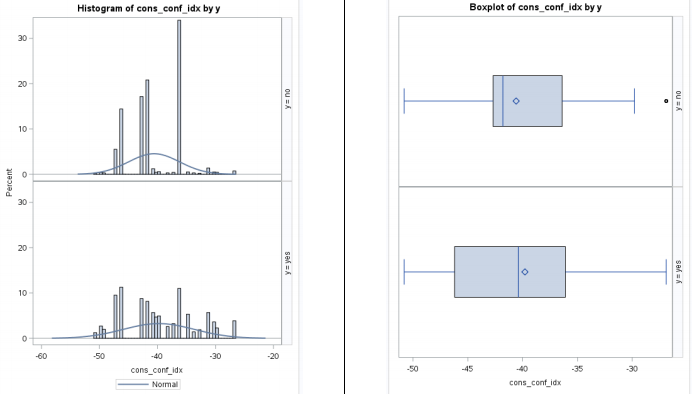
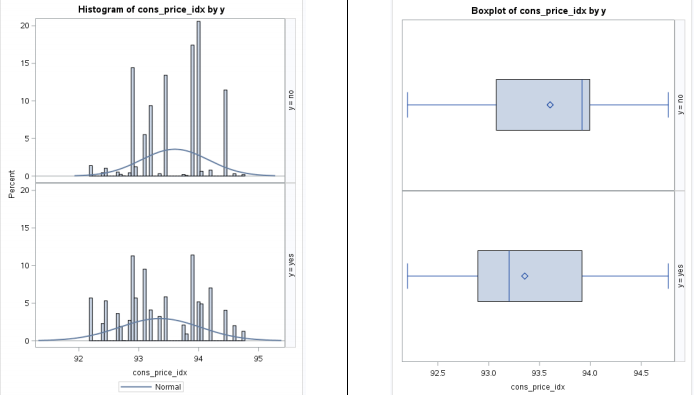
No hay missings en este dataset. Un vistazo a los gráficos revela que la edad del cliente no tiene influencia en la variable objetivo (y); los histogramas estan centrados en la misma región y tienen varianza parecida sin importar el valor de y.

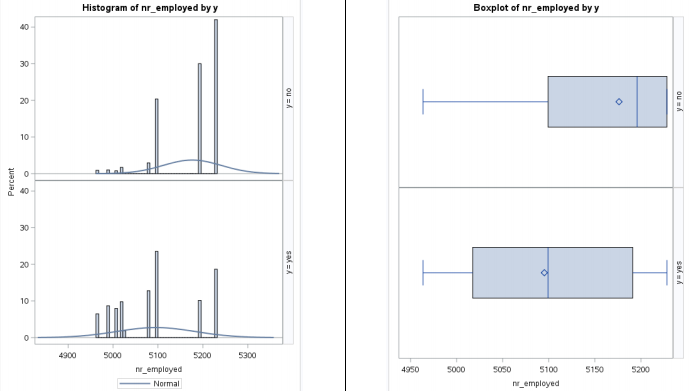
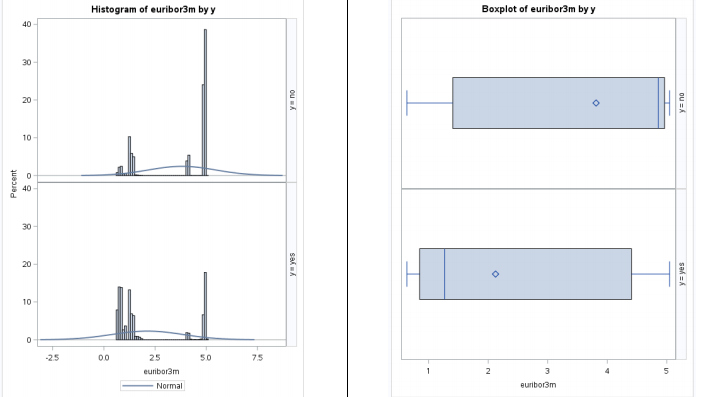
Las siguientes clasificaciones se hacen en base a los histogramas, diagramas de cjas y frecuencias de cada varible pues estos revelan que pueden ser agrupadas. Además, esto será más económico computacionalmente.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Continuous** |  |  |  | **Figure** |  |  | **Figure Notes** |  |  | **Categorization Notes** |  |  | **Categorized** |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  | **Variable** |  |  |  |  |  |  |  |  |  |  |  |  | **Variable** |  |
|  | campaign |  |  | | 1 |  |  | > 97% de los datos bajo los10 |  |  | Ordinal en {1,2,3,>3) |  |  | campaign\_cat |  |
|  |  |  |  |  |  |  |  | niveles para ambos niveles |  |  |  |  |  |  |  |
|  |  |  |  |  | |  |  | | --- | --- | | cons\_price\_idx\_cat |  | |  |  | de respuesta. |  |  |  |  |  |  |  |
|  | pdays | | 2 | | |  |  | > 79% del dato pdays = |  |  | Binario en {‘contacted before’ y |  |  | pdays\_cat | |
|  |  |  |  |  |  |  |  | 999 (no contacto previo) |  |  | ‘never contacted’} |  |  |  |  |
|  |  |  |  |  |  |  |  | Para los dos niveles de |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | respuesta. |  |  |  |  |  |  |  |
|  | previous |  |  | | 3 |  |  | >68% del dato en los nive- |  |  | Binarios en {‘contacted before’ y |  |  | previous\_cat |  |
|  |  |  |  |  |  |  |  | les mas pequeños para |  |  | ‘never contacted’} |  |  |  |  |
|  |  |  |  |  |  |  |  | Ambos niveles. |  |  |  |  |  |  |  |
|  | emp\_var\_rate | | 4 | | |  |  | Por los graficos de barras, |  |  | Ordinal en {<=-1.8, (-1.8 to -0.1], > - |  |  | emp\_var\_rate\_cat | |
|  |  |  |  |  |  |  |  | La clasificación parece{<=- |  | 0.1} | |  |  |  |  |
|  |  |  |  |  |  |  |  | 1.8, (-1.8 to -0.1], > -0.1} |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | para ambos niveles. |  |  |  |  |  |  |  |
|  | cons\_price\_idx |  |  |  | 5 |  |  | Datos multimodales agrupados en cuartiles. |  |  |  |  |  | cons\_price\_idx\_cat |  |
|  | cons\_conf\_idx |  |  |  | 6 |  |  | Datos multimodales organizados en sixtiles. |  |  |  |  |  | cons\_conf\_idx\_cat |  |
|  |  | |  | | |  |  |  |  |  |  |  |  |  | |
|  | euribor3m |  |  |  | 7 |  |  | Datos multimodales organizados en sixtiles. |  |  |  |  |  | euribor3m\_cat |  |
|  | nr\_employed |  |  |  | 8 |  |  | Datos multimodales organizados en cuartiles. |  |  |  |  |  | nr\_employed\_cat |  |





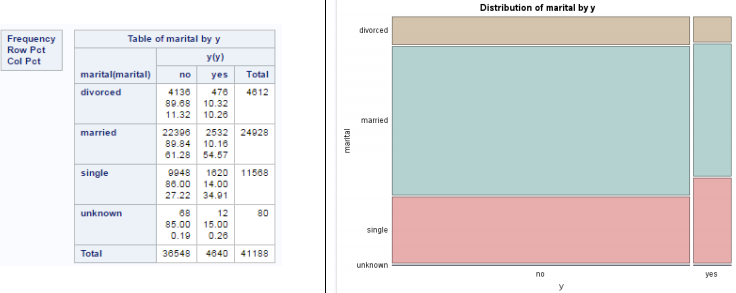
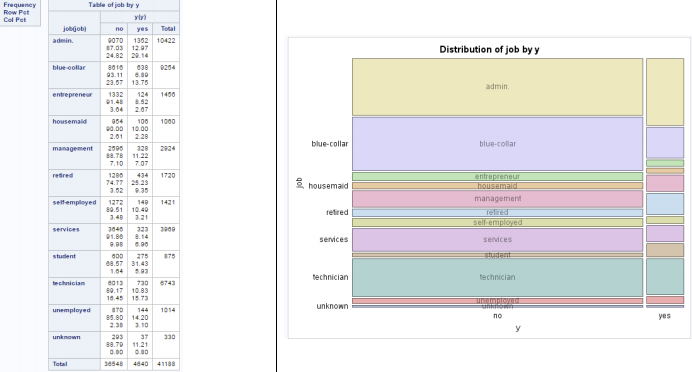


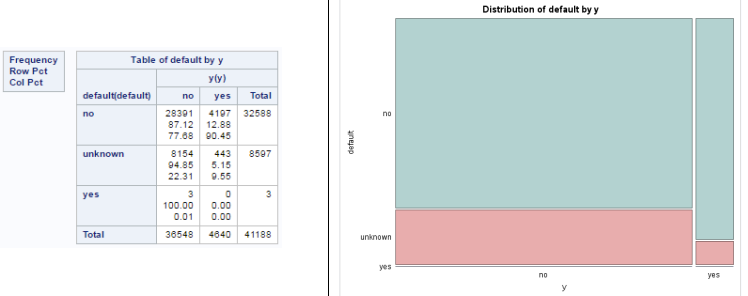
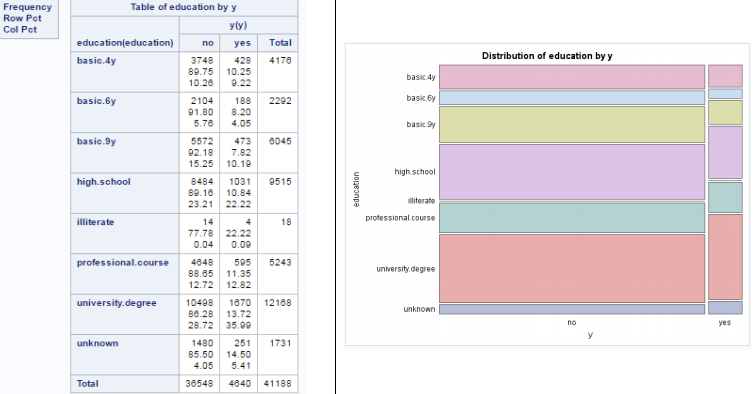


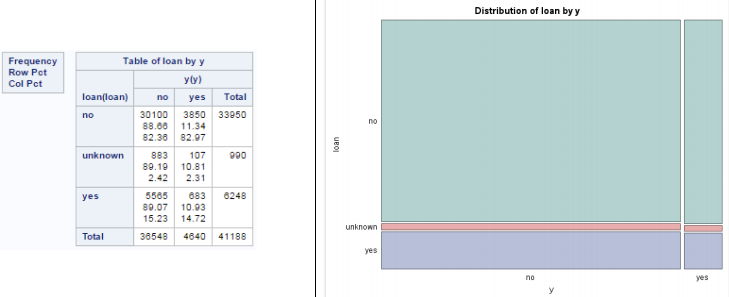
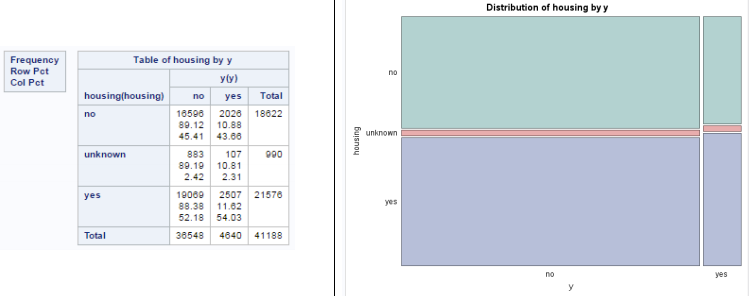
4.2. categóricas

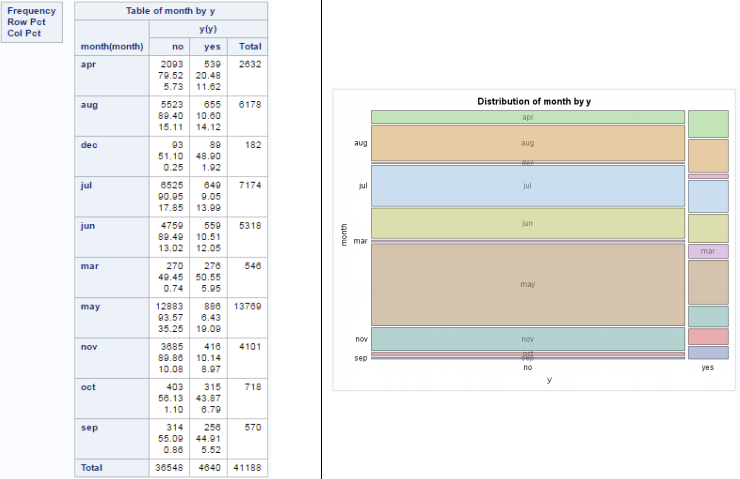
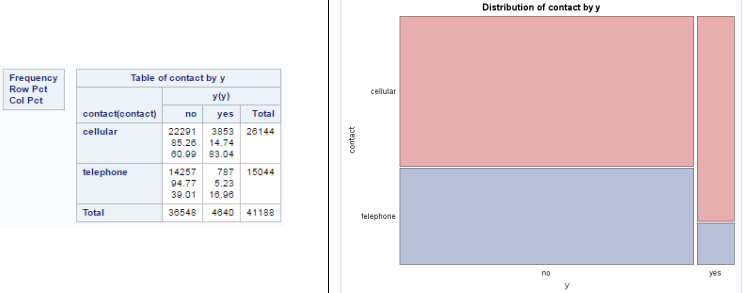
No hay missing values. Puesto que; campaign, previous, emp\_var\_rate, cons\_price\_idx, cons\_conf\_idx, euribor3m y nr\_employed fueron codificadas como variables categóricas, se analizarán aquí también. Una revisión de los gráficos de mosaico y las frecuencias revela lo siguiente:

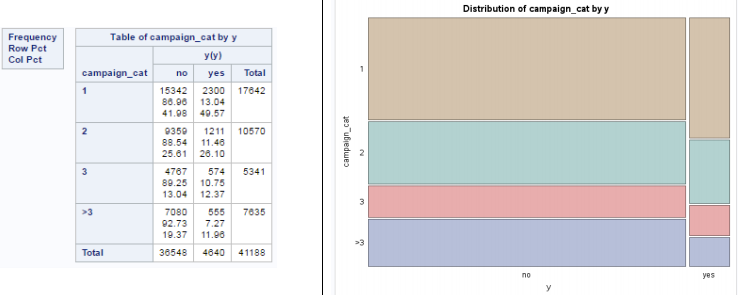
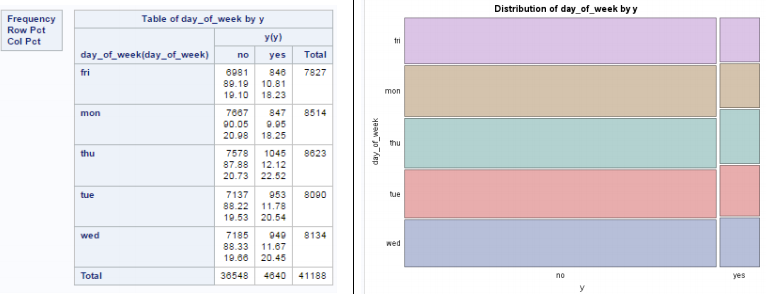
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Variable categórica** |  |  | **Notas de la figura** |  |
|  |  |  |  |
|  | job |  |  | Nivel 'desconocido', 1.6% de observaciones, se añade a la categoría más grande de 'admin'. |  |
|  |  |  |  |  |  |
|  | marital |  |  | Nivel 'desconocido',. 45% del total de las observaciones, será lanzado en el categoría más grande de 'casado'. | |
|  |  |  |  |  | |
|  | Education |  |  | Nivel 'desconocido’ , 9.46% de las observaciones, se añade en la categoría más grande de 'university.degree'. También, puesto que sólo hay 18 observaciones |  |
|  |  |  |  | para ‘illiterate’, esta categoria será eliminada. 18 |  |
|  |  |  |  | observaciones no es suficiente para hacer una inferencia adecuada. |  |
|  |  |  |  |  |  |
|  | default |  |  | Nivel 'sí',. 01% de las observaciones, se eliminará. 3 observaciones no es suficiente para hacer una inferencia adecuada. | |
|  |  |  |  |  | |
|  | housing |  |  | Nivel 'desconocido', 4,73% del total de las observaciones, será añadida a la categoría más grande del 'sí'. |  |
|  |  |  |  |  |  |
|  | loan |  |  | Nivel 'desconocido', 4,73% de las observaciones, será añadida a la categoría más grande del 'no'. | |
|  |  |  |  |  | |
|  | contact, month, |  |  | Sin variar. |  |
|  | campaign\_cat, |  |  |  |  |
|  | previous\_cat, |  |  |  |  |
|  | poutcome, |  |  |  |  |
|  | emp\_var\_rate\_cat, |  |  |  |  |
|  | cons\_price\_idx\_cat, |  |  |  |  |
|  | cons\_conf\_idx\_cat, |  |  |  |  |
|  | euribor3m, |  |  |  |  |
|  | nr\_employed\_cat |  |  |  |  |
|  | day\_of\_week |  |  | Sin variar. | |

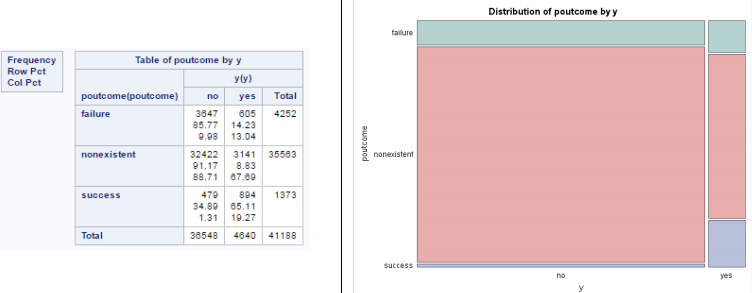
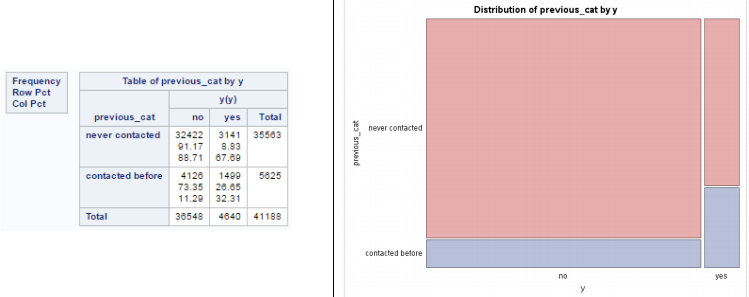


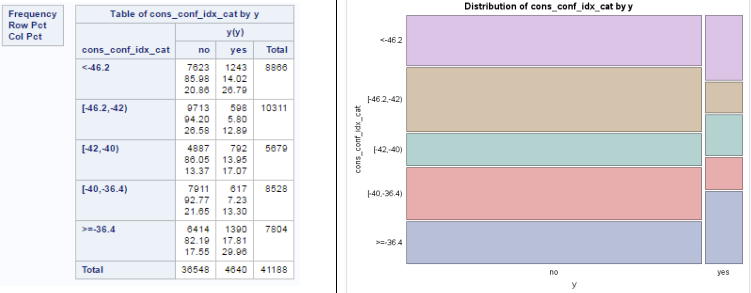
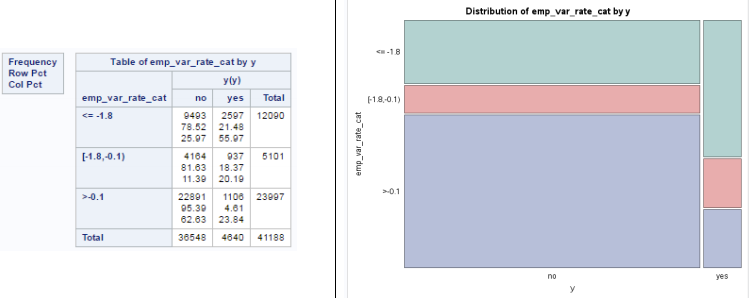


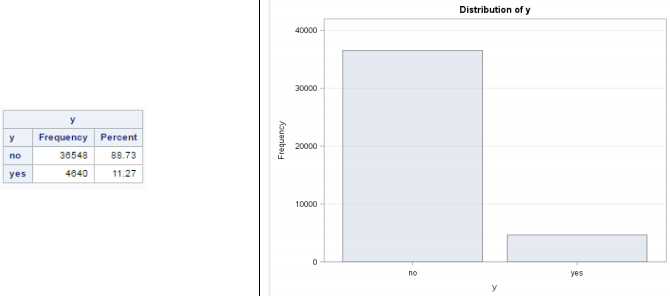
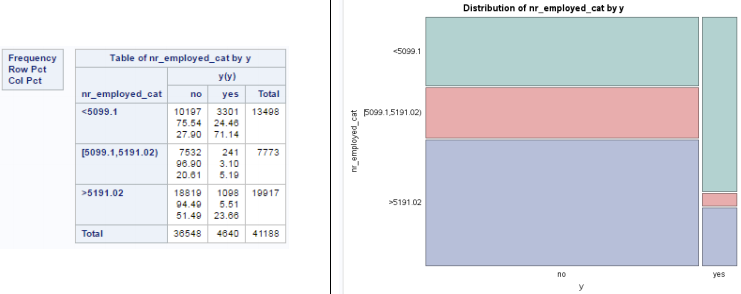
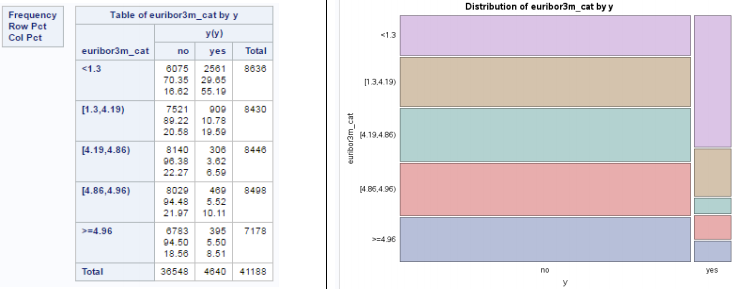












5. Análisis

Mirando los gráficos de mosaico ya nos podemos hacer a la idea de que variables influyen en la variable target (y) y cuales no. Una distribución parecida en ‘si’ y ‘no’ de las diferentes categorias significa que no son significativas. Por otro lado, cuanto más desiguales sean las distribuciones en las dos etiquetas más significativas serán las variables a la hora de predecir.

Puesto que sólo el 11,27% de los datos tiene respuestas 'sí' para la variable target, los modelos se construirán en un conjunto de datos compuesto por todas las respuestas 'sí' (4.636) y una muestra aleatoria de 4.636 respuestas 'no' (\*).  Esto ayudará a que podamos ejecutar una macro con todas las variables puesto que con un dataset mayor no se podria ejecutar con la capa gratuita de sas.

A continuación, los datos se dividen 50/50 en conjuntos de datos de entrenamiento y validación (\*\*).

\*La muestra aleatoria de los 'no' se estratifica por job, civil, education, default, housing, loan y month para asegurar que la muestra se asemeja al dataset principal lo máximo posible. Este método da el modelo de mayor calidad para detectar qué variables impactan respuesta objetivo (y).

\*\* Los conjuntos de datos de entrenamiento y validación serán estratificados por education y job, ya que estas 2 variables tienen la mayoría de las categorías. Esto afirma que se pueden hacer inferencias adecuadas de cualquier conjunto de datos.

proc sql noprint;

create table banco\_yes as select \* from banco03 where(y EQ 1);

quit;

proc sql noprint;

create table banco\_no as select \* from banco03 where(y EQ 0);

quit;

proc sort data=banco\_no out=WORK.SORTTempTableSorted;

by job marital education default housing loan month;

run;

proc surveyselect data=WORK.SORTTempTableSorted out=BANCO\_SAMPLE\_NO

method=srs sampsize=4636;

strata job marital education default housing loan month / alloc=prop;

run;

proc delete data=WORK.SORTTempTableSorted;

run;

data banco\_modeling;

set banco\_yes banco\_sample\_no;

run;

proc sort data=BANCO\_MODELING out=work.\_sorted\_;

by education job;

run;

proc means data=work.\_sorted\_ noprint;

by education job;

output out=work.\_meansOut\_(drop=\_type\_ \_freq\_) n=\_\_nobs\_\_;

run;

proc sql noprint;

select max(\_\_nobs\_\_) into :count from work.\_meansOut\_;

quit;

data banco\_train banco\_validate;

set work.\_sorted\_;

by education job;

retain \_\_tmp1-\_\_tmp%trim(&count) \_\_nobs\_\_ \_\_nobs1\_\_ \_\_nobs2\_\_;

retain \_\_nobs\_\_ \_\_seed\_\_ \_n1\_;

drop \_\_k\_\_;

drop \_i\_ \_\_seed\_\_ \_\_tmp1-\_\_tmp%trim(&count);

drop \_n1\_ \_\_nobs\_\_ \_\_nobs1\_\_ \_\_nobs2\_\_;

array \_\_tmp(\*) \_\_tmp1-\_\_tmp%trim(&count);

if (\_n\_=1) then

do;

\_\_seed\_\_=9889;

\_\_nobs\_\_=&count;

end;

if first.job then

do;

set work.\_meansOut\_;

by education job;

do \_i\_=1 to \_\_nobs\_\_;

\_\_tmp(\_i\_)=\_i\_;

end;

if (\_\_nobs\_\_ < dim(\_\_tmp)) then

do;

do \_i\_=\_\_nobs\_\_+1 to dim(\_\_tmp);

\_\_tmp(\_i\_)=0; end;

end;

call ranperm(\_\_seed\_\_, of \_\_tmp(\*));

if (\_\_nobs\_\_ < dim(\_\_tmp)) then do;

\* mover los valores 0 al comienzo de la lista;

do \_i\_=1 to dim(\_\_tmp);

if (\_\_tmp(\_i\_)=0) then

do;

if (\_i\_ < dim(\_\_tmp)) then

do;

\_\_k\_\_=\_i\_ + 1;

do while(\_\_k\_\_ < dim(\_\_tmp)

and \_\_tmp(\_\_k\_\_)=0);

\_\_k\_\_=\_\_k\_\_+1;

end;

if (\_\_k\_\_ <=dim(\_\_tmp))

then do;

\_\_tmp(\_i\_)=\_\_tmp(\_\_k\_\_);

\_\_tmp(\_\_k\_\_)=0; end; end; end; end; end;

\_n1\_=0;

\_\_nobs1\_\_=round(0.5\*\_\_nobs\_\_);

\_\_nobs2\_\_=round(0.5\*\_\_nobs\_\_)+\_\_nobs1\_\_; end;

\_n1\_=\_n1\_ + 1;

if (\_n1\_ <=dim(\_\_tmp)) then do;

if (\_\_tmp(\_n1\_) > 0) then do;

if (\_\_tmp(\_n1\_) <=\_\_nobs1\_\_) then do;

output banco\_train; end;

else if (\_\_tmp(\_n1\_) <=\_\_nobs2\_\_) then do; output banco\_validate;end; end; end; run;

proc delete data=work.\_sorted\_; run;

proc delete data=work.\_meansOut\_; run;

6. Resultados/generalización

6.1. Selección características principales

Empleo la siguiente macro para ver que variables son las más significativas a la hora de predecir la varible target (y) en una regresión logística(\*).

%macro logistic (t\_input, vardepen, varindep, interaccion, semi\_ini, semi\_fin );

ods trace on /listing;

%do semilla=&semi\_ini. %to &semi\_fin.;

ods output EffectInModel= efectoslog;/\*Test de Wald de efectos en el modelo\*/

ods output FitStatistics= ajustelog; /\*"Estadisticos de ajuste", AIC \*/

ods output ParameterEstimates= estimalog;/\*"Estimadores de parametro"\*/

ods output ModelBuildingSummary=modelolog; /\*Resumen modelo, efectos\*/

ods output RSquare=ajusteRlog; /\*R-cuadrado y Max-rescalado R-cuadrado\*/

proc logistic data=&t\_input. EXACTOPTIONS (seed=&semilla.) ;

class &varindep.;

model &vardepen. = &varindep. &interaccion.

/ selection=stepwise details rsquare NOCHECK;

run;

data un1; i=12; set efectoslog; set ajustelog; point=i; run;

data un2; i=12; set un1; set estimalog; point=i; run;

data un3; i=12; set un2; set modelolog; point=i; run;

data union&semilla.; i=12; set un3; set ajusteRlog; point=i; run;

proc append base=t\_models data=union&semilla. force; run;

proc sql; drop table union&semilla.; quit;

%end;

ods html close;

proc sql; drop table efectoslog,ajustelog,ajusteRlog,estimalog,modelolog; quit;

%mend;

%logistic (banco\_train, y , age job marital education 'default'n housing loan contact 'month'n day\_of\_week campaign

pdays previous poutcome 'emp.var.rate'n 'cons.price.idx'n 'cons.conf.idx'n euribor3m

'nr.employed'n pdays\_cat campaign\_cat previous\_cat emp\_var\_rate\_cat cons\_price\_idx\_cat

cons\_conf\_idx\_cat euribor3m\_cat nr\_employed\_cat , age job marital education 'default'n housing loan contact 'month'n day\_of\_week campaign pdays previous poutcome 'emp.var.rate'n 'cons.price.idx'n 'cons.conf.idx'n euribor3m

'nr.employed'n pdays\_cat campaign\_cat previous\_cat emp\_var\_rate\_cat cons\_price\_idx\_cat

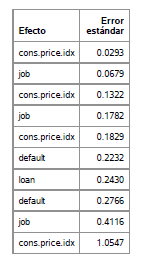
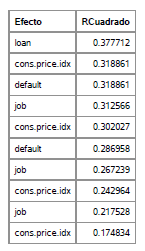
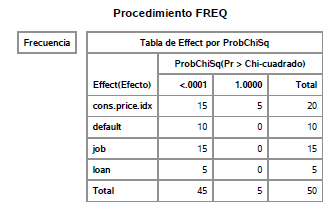
cons\_conf\_idx\_cat euribor3m\_cat nr\_employed\_cat , 12345, 12349);

/\*Analisis de los resultados obtenidos de la macro\*/

proc freq data=t\_models (keep=effect ProbChiSq); tables effect\*ProbChiSq /norow nocol nopercent; run;

proc sql; select distinct \* from t\_models (keep=effect nvalue1 rename=(nvalue1=RCuadrado))

El resultado es el siguiente:

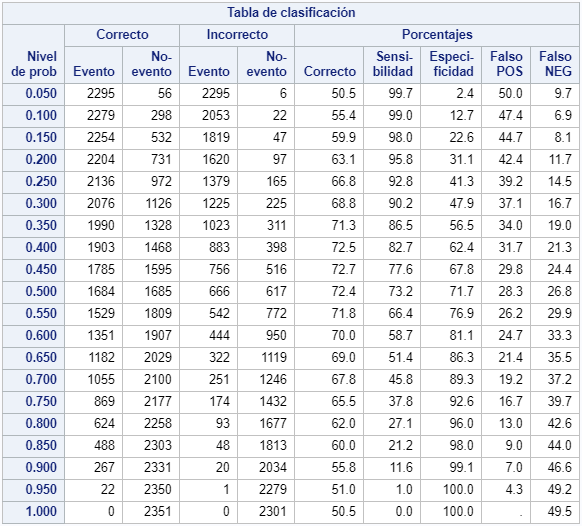
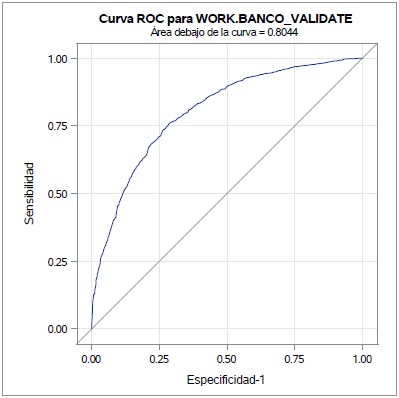
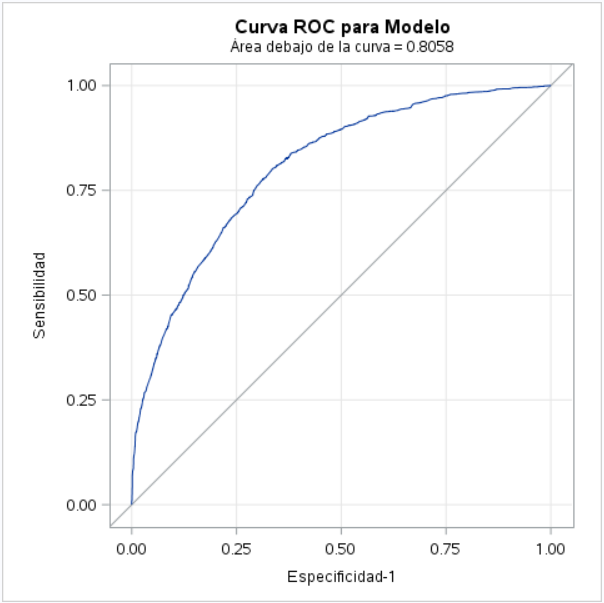


Cabe destacar que los Rcuadrado no parecen gran cosa ya que son lejanos a 1, pero es lo máximo que he encontrado probando a: combinar variables y subir el número de semillas.

\* Las variables con un error estándar mayor que 1 o entrarán en el modelo sean o no significativas.

6.2. Conclusiones de la regresión logística

Entreno un modelo de regresión logística con las caracteristicas seleccionadas por el modelo ('cons.price.idx'n default job loan) y visualizo su curva roc tanto en training como en validacion para comprobar que la selección de características es optima.



ods graphics on;

ods output CLOddsPL=banco\_OddsRatiosPL;

proc logistic data=banco\_train;

class y(desc) 'cons.price.idx'n default job loan

/param=ref;

model y(event="1") = 'cons.price.idx'n default job loan / outroc=banco\_troc

cl clodds=both clparm=both ctable plcl ;

score data=banco\_validate out=banco\_valpred outroc=banco\_vroc fitstat;

roc; roccontrast;

run;

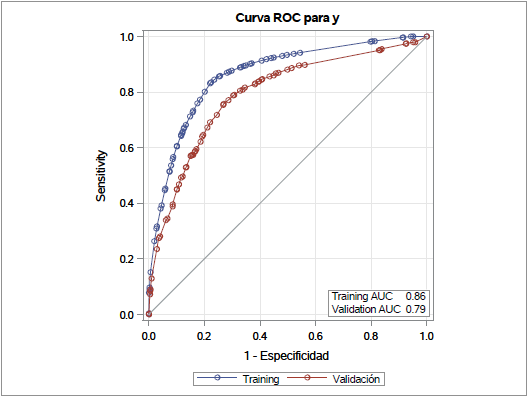
A pesar de tener un Rcuadrado mediocre, el clasificador compuesto por estas variables parece ser decente con un AUC de 0.8 tanto para training como para validación por lo quje decido quedarme con ese conjunto de caracteristicas para el modelo de regresión logística.

6.3. Otros modelos y comparación con artículo de sas

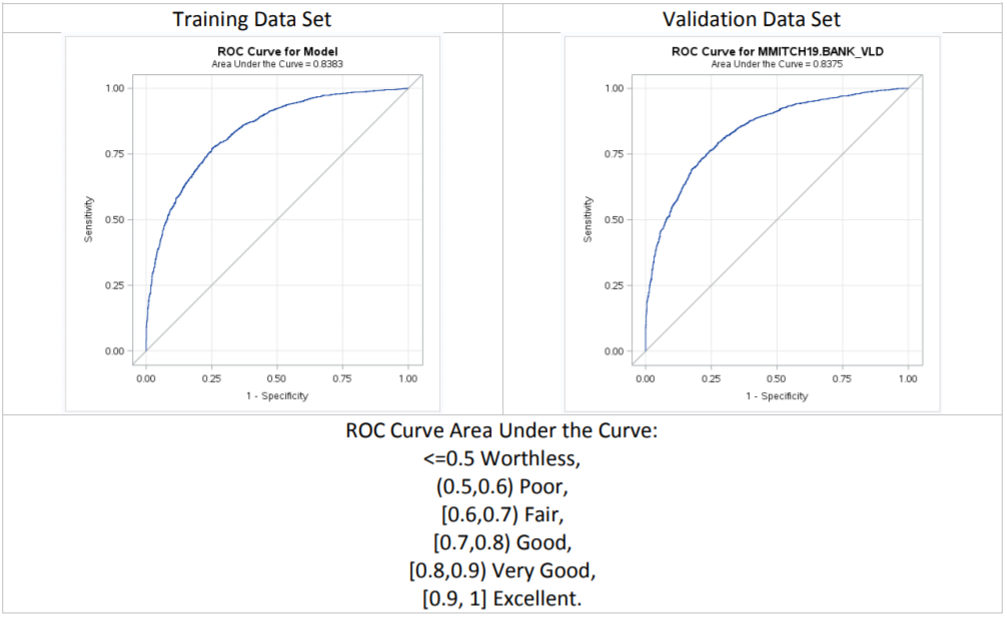
En primer lugar, he seguido el preprocesamiento que hacen porque me parece el mejor que se podría hacer a este dataset en concreto. Además los tipos de gráficos que sugieren tambien me lo parecen.

En segundo lugar, ellos escogen las variables haciendo un ranking de los AUC de cada regresión logística de la forma y=f(característica). De la siguiente manera escogen las variables de cada tipo (datos de los clientes, ultima información del contacto, otros, variables sociales y económicas) que mejor modelicen la variable target. De esta manera, ellos obtienen(\*) una AUC de 0.83 (con: job marital education default housing loan contact month day\_of\_week campaign\_cat previous\_cat poutcome emp\_var\_rate\_cat cons\_price\_idx\_cat cons\_conf\_idx\_cat euribor3m\_cat nr\_employed\_cat) mientra que nosotros obtenemos un AUC del 0.8 (con: 'cons.price.idx'n default job loan).

En tercer lugar, teniendo en cuenta que la variable target es dicotómica, no seria lógico emplear una regresión lineal. Más sentido tendría emplear arboles de decisión u otros algoritmos enfocados a clasificar. En cuanto a la red neuronal que se sugiere hacer he de decir que me he tomado tiempo en intentar hacer una clasificación empleando diferentes arquitecturar y una función de clasificación softmax pero no he conseguido hacer que funcionara del todo bien. Cabe mencionar también que he empleado un árbol de decisión como el que emplean ellos obteniendo unas prestaciones similares a las que obtiene el modelo logístico para training aunque tiende al overfitting:

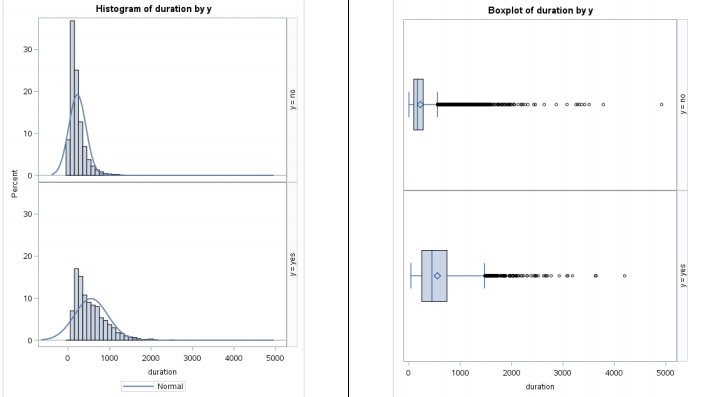
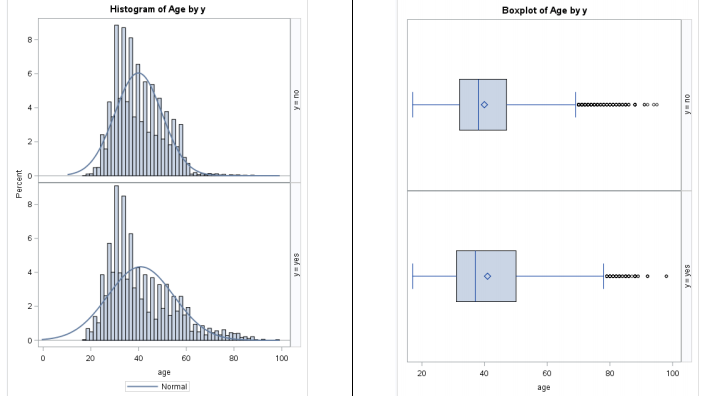


(\*)



9. Apéndice: Tablas, gráficos y código SAS

* Gráficos de las variables no tenidas en cuenta:

* Código empleado en toda la practica:

libname datos '/home/miguelcorgen0/my\_courses/cores/my\_project';

data banco01;

set datos.bank\_additional\_full;

run;

\* ANALISIS EXPLORATORIO ;

PROC CONTENTS DATA=banco01; RUN;

\* Hay NaS????;

PROC MI Data=banco01 simple; run;

\* Histogramas y graficos de cajas de las variables continuas;

proc sgpanel data=banco01;

title "Histograma age - y";

panelby y / layout=rowlattice;

histogram age;

density age;

run;

proc sgpanel data=banco01;

title "Boxplot age - y";

panelby y / layout=rowlattice;

hbox age;

run;

proc sgpanel data=banco01;

title "Histograma duration - y";

panelby y / layout=rowlattice;

histogram duration;

density duration;

run;

proc sgpanel data=banco01;

title "Boxplot of duration by y";

panelby y / layout=rowlattice;

hbox duration;

run;

proc sgpanel data=banco01;

title "Histogram of campaign by y";

panelby y / layout=rowlattice;

histogram campaign;

density campaign;

run;

proc sgpanel data=banco01;

title "Boxplot of campaign by y";

panelby y / layout=rowlattice;

hbox campaign;

run;

proc sgpanel data=banco01;

title "Histogram of pdays by y";

panelby y / layout=rowlattice;

histogram pdays;

density pdays;

run;

proc sgpanel data=banco01;

title "Boxplot of pdays by y";

panelby y / layout=rowlattice;

hbox pdays;

run;

proc sgpanel data=banco01;

title "Histogram of previous by y";

panelby y / layout=rowlattice;

histogram previous;

density previous;

run;

proc sgpanel data=banco01;

title "Boxplot of previous by y";

panelby y / layout=rowlattice;

hbox previous;

run;

proc sgpanel data=banco01;

title "Histogram of emp\_var\_rate by y";

panelby y / layout=rowlattice;

histogram 'emp.var.rate'n;

density 'emp.var.rate'n;

run;

proc sgpanel data=banco01;

title "Boxplot of emp\_var\_rate by y";

panelby y / layout=rowlattice;

hbox 'emp.var.rate'n;

run;

proc sgpanel data=banco01;

title "Histogram of cons\_price\_idx by y";

panelby y / layout=rowlattice;

histogram 'cons.price.idx'n;

density 'cons.price.idx'n;

run;

proc sgpanel data=banco01;

title "Boxplot of cons\_price\_idx by y";

panelby y / layout=rowlattice;

hbox 'cons.price.idx'n;

run;

proc sgpanel data=banco01;

title "Histogram of cons\_conf\_idx by y";

panelby y / layout=rowlattice;

histogram 'cons.conf.idx'n ;

density 'cons.conf.idx'n ;

run;

proc sgpanel data=banco01;

title "Boxplot of cons\_conf\_idx by y";

panelby y / layout=rowlattice;

hbox 'cons.conf.idx'n ;

run;

proc sgpanel data=banco01;

title "Histogram of euribor3m by y";

panelby y / layout=rowlattice;

histogram euribor3m;

density euribor3m;

run;

proc sgpanel data=banco01;

title "Boxplot of euribor3m by y";

panelby y / layout=rowlattice;

hbox euribor3m;

run;

proc sgpanel data=banco01;

title "Histogram of nr\_employed by y";

panelby y / layout=rowlattice;

histogram 'nr.employed'n;

density 'nr.employed'n;

run;

proc sgpanel data=banco01;

title "Boxplot of nr\_employed by y";

panelby y / layout=rowlattice;

hbox 'nr.employed'n;

run;

options validvarname=any;

proc format;

value campaign 1 = '1' 2 = '2' 3='3' 4='>3';

value previous 0 = 'never contacted' 1='contacted before';

value emp\_var\_rate 1 = '<= -1.8' 2='[-1.8,-0.1)' 3='>-0.1';

value cons\_price\_idx 1 = '<93.06' 2='[93.06,93.91]' 3='>93.91';

value cons\_conf\_idx 1='<-46.2' 2='[-46.2,-42)' 3='[-42,-40)' 4='[-40,-36.4)' 5='>=-36.4';

value euribor3m 1='<1.3' 2='[1.3,4.19)' 3='[4.19,4.86)' 4='[4.86,4.96)' 5='>=4.96';

value nr\_employed 1 = '<5099.1' 2='[5099.1,5191.02)' 3='>5191.02';

\* Frequency Table of campaign for recoding;

proc sort data=banco01 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted order=freq;

tables campaign / plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

\* Frequency Table of pdays;

proc sort data=banco01 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted order=freq;

tables pdays / missing plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

\* Frequency Table of previous;

proc sort data=banco01 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted order=freq;

tables previous / missing plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

/\*--Set output size--\*/

ods graphics / reset imagemap;

/\*--SGPLOT proc statement--\*/

proc sgplot data=banco01;

/\*--TITLE and FOOTNOTE--\*/

title 'Grouped Bar Chart of emp\_var\_rate by y';

/\*--Bar chart settings--\*/

vbar 'emp.var.rate'n / group=y groupdisplay=Cluster name='Bar';

/\*--Response Axis--\*/

yaxis grid;

run;

ods graphics / reset;

title;

\* Frequency Table of cons\_price\_idx;

proc sort data=banco01 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted order=freq;

tables 'cons.price.idx'n / missing plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

/\*--Set output size--\*/

ods graphics / reset imagemap;

/\*--SGPLOT proc statement--\*/

proc sgplot data=banco01;

/\*--TITLE and FOOTNOTE--\*/

title 'Grouped Bar Chart of cons\_price\_idx by y';

/\*--Bar chart settings--\*/

vbar 'cons.price.idx'n / group=y groupdisplay=Cluster name='Bar';

/\*--Response Axis--\*/

yaxis grid;

run;

ods graphics / reset;

title;

ods noproctitle;

\*Bucket binning for cons\_price\_idx;

proc hpbin data=banco01 numbin=3 bucket computestats computequantile;

input 'cons.price.idx'n;

run;

\* Frequency Table of cons\_conf\_idx;

proc sort data=banco01 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted order=freq;

tables 'cons.conf.idx'n / missing plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

\*Set output size ods graphics / reset imagemap;

\*SGPLOT proc statement;

proc sgplot data=banco01;

\*TITLE and FOOTNOTE;

title 'Grouped Bar Chart of cons\_conf\_idx by y';

\*Bar chart settings;

vbar 'cons.conf.idx'n / group=y groupdisplay=Cluster name='Bar';

\*Response Axis;

yaxis grid;

run;

ods graphics / reset;

title;

\*Quantile binning for cons\_conf\_idx;

ods noproctitle;

proc hpbin data=banco01 numbin=5 pseudo\_quantile computestats

computequantile;

input 'cons.conf.idx'n;

run;

ods noproctitle;

\*Quantile binning for euribor3m;

proc hpbin data=banco01 numbin=5 pseudo\_quantile computestats

computequantile;

input euribor3m;

run;

\*Quantile binning for nr\_employed;

ods noproctitle;

proc hpbin data=banco01 numbin=4 pseudo\_quantile;

input 'nr.employed'n;

run;

\*Creating categorical variables for continuous variables that require them;

data banco02;

set banco01;

pdays\_cat = 'never contacted';

if pdays ^= 999 then pdays\_cat = 'contacted before';

campaign\_cat = 999;

if campaign = 1 then campaign\_cat = 1;

if campaign = 2 then campaign\_cat = 2;

if campaign = 3 then campaign\_cat = 3;

if campaign > 3 then campaign\_cat = 4;

previous\_cat = 1;

if previous = 0 then previous\_cat = 0;

emp\_var\_rate\_cat = 999;

if 'emp.var.rate'n LE -1.8 then emp\_var\_rate\_cat = 1;

if ('emp.var.rate'n > -1.8) and ('emp.var.rate'n LE -0.1) then emp\_var\_rate\_cat = 2;

if 'emp.var.rate'n > -0.1 then emp\_var\_rate\_cat = 3;

cons\_price\_idx\_cat = 999;

if 'cons.price.idx'n < 93.056333333 then cons\_price\_idx\_cat = 1;

if ('cons.price.idx'n GE 93.056333333) and ('cons.price.idx'n LE 93.911666667) then

cons\_price\_idx\_cat = 2;

if 'cons.price.idx'n > 93.911666667 then cons\_price\_idx\_cat = 3;

cons\_conf\_idx\_cat = 999;

if 'cons.conf.idx'n < -46.19925 then cons\_conf\_idx\_cat = 1;

if ('cons.conf.idx'n GE -46.19925) and ('cons.conf.idx'n LE -41.99763) then

cons\_conf\_idx\_cat = 2;

if ('cons.conf.idx'n GE -41.99763) and ('cons.conf.idx'n LE -39.99959) then

cons\_conf\_idx\_cat = 3;

if ('cons.conf.idx'n GE -39.99959) and ('cons.conf.idx'n LE -36.39786) then

cons\_conf\_idx\_cat = 4;

if 'cons.conf.idx'n > -36.39786 then cons\_conf\_idx\_cat = 5;

euribor3m\_cat = 999;

if euribor3m < 1.2991788 then euribor3m\_cat = 1;

if (euribor3m GE 1.2991788) and (euribor3m LE 4.1910304) then euribor3m\_cat = 2;

if (euribor3m GE 4.1910304) and (euribor3m LE 4.864149) then euribor3m\_cat = 3;

if (euribor3m GE 4.864149) and (euribor3m LE 4.9620732) then euribor3m\_cat = 4;

if euribor3m > 4.9620732 then euribor3m\_cat = 5;

nr\_employed\_cat = 999;

if 'nr.employed'n < 5099.10335 then nr\_employed\_cat = 1;

if ('nr.employed'n GE 5099.10335) and ('nr.employed'n LE 5191.0171) then

nr\_employed\_cat = 2;

if 'nr.employed'n > 5191.0171 then nr\_employed\_cat = 3;

run;

/\*Frequency table for emp\_var\_rate\_cat\*/;

proc sort data=banco02 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted;

format emp\_var\_rate\_cat emp\_var\_rate.;

tables emp\_var\_rate\_cat / plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

/\*Frequency table for cons\_price\_idx\_cat\*/;

proc sort data=banco02 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted;

format cons\_price\_idx\_cat cons\_price\_idx.;

tables cons\_price\_idx\_cat / plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

/\*Frequency table for cons\_conf\_idx\_cat\*/;

proc sort data=banco02 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted;

format cons\_conf\_idx\_cat cons\_conf\_idx.;

tables cons\_conf\_idx\_cat / plots=none; by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

/\*Frequency table for euribor3m\_cat \*/;

proc sort data=banco02 out=Work.SortTempTableSorted; by y;

run;

proc freq data=Work.SortTempTableSorted;

format euribor3m\_cat euribor3m.;

tables euribor3m\_cat / plots=none;

by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

/\*Frequency table for nr\_employed\_cat\*/;

proc sort data=banco02 out=Work.SortTempTableSorted;

by y;

run;

proc freq data=Work.SortTempTableSorted;

format nr\_employed\_cat nr\_employed.;

tables nr\_employed\_cat / plots=none;

by y;

run;

proc delete data=Work.SortTempTableSorted;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Categorical Variables\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

ods noproctitle;

proc freq data=banco02;

tables (job) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (marital) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (education) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (default) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (housing) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (loan) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (contact) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (month) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (day\_of\_week) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format campaign\_cat campaign.;

tables (campaign\_cat) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format previous\_cat previous.;

tables (previous\_cat) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (poutcome) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format emp\_var\_rate\_cat emp\_var\_rate.;

tables (emp\_var\_rate\_cat) \*(y) / missing nopercent nocum

plots(only)=(freqplot mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format cons\_price\_idx\_cat cons\_price\_idx.;

tables (cons\_price\_idx\_cat) \*(y) / missing nopercent nocum

plots(only)=(freqplot mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format cons\_conf\_idx\_cat cons\_conf\_idx.;

tables (cons\_conf\_idx\_cat) \*(y) / missing nopercent nocum

plots(only)=(freqplot mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format euribor3m\_cat euribor3m.;

tables (euribor3m\_cat) \*(y) / missing nopercent nocum plots(only)=(freqplot

mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

format nr\_employed\_cat nr\_employed.;

tables (nr\_employed\_cat) \*(y) / missing nopercent nocum

plots(only)=(freqplot mosaicplot);

run;

ods noproctitle;

proc freq data=banco02;

tables (y) / missing nocum plots(only)=(freqplot mosaicplot);

run;

/\*Collapsing categorical variables unknown level and levels that can be

collapsed; Recode y into 1=yes, 0=no\*/;

data banco03;

set banco02;

if job = 'unknown' then job = 'admin.';

if marital = 'unknown' then marital = 'married';

if education = 'unknown' then education = 'university.degree';

if education = 'illiterate' then DELETE;

if default = 'yes' then DELETE;

if housing = 'unknown' then housing = 'yes';

if loan = 'unknown' then loan = 'no';

if y = 'yes' then y2 =1;

if y = 'no' then y2=0;

drop y;

rename y2=y;

run;

ods graphics on;

proc hpsplit data=banco\_modeling maxdepth=9;

class y 'cons.price.idx'n default job loan;

model y(event="1") = 'cons.price.idx'n default job loan;

prune costcomplexity; \*(leaves=15) ;

partition fraction(validate = 0.5 seed=9889) ;

run;