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. 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 - collar', 'emprendedor', 'criada', 'gestión', 'jubilado',' auto - empleados ',' servicios 'estudiante', 'técnico', 'parado', 'desconocido')	Categóricos
marital	Datos del cliente	Estado civil de clientes en el momento de la llamada - 'divorciado', 'casado', 'solo', 'desconocido'; Nota: significa 'divorciada' divorciado o viudo	Categóricos
Education	Datos del cliente	Formación de clientes en el momento de la llamada- ' basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'	Categóricos
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

Ioan	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 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.	Continua
campaign	Otros	Número de contactos realizados durante esta campaña para este cliente (incluye ultima en contacto con)	Continua
pdays	Otros	Número de días que pasa después de pasado el cliente contactado de una anterior campaña (numérico; 999 significa cliente no contactó previamente)	Continua
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 mide los cambios en los precios pagados por los consumidores por una canasta de bienes y servicios de cada mes.	Continua
cons_conf_idx	Social y económico	Índice de confianza del consumidor – indicador mensual; En Portugal, el consumidor 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.	Continua
nr_employed	Social y económico	Euribor 3 meses tasa – Indicador diario; Euribor es corto para Euro Interbank 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.	Continua
Nr.employed	Social y económico	Número de empleados – indicador trimestral; Número de personas empleadas para un cuarto.	Continua
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

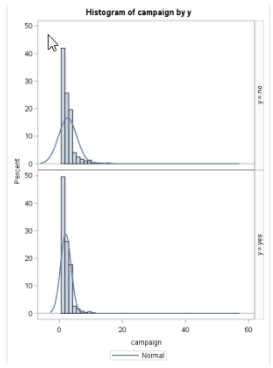
4.1. Variables continuas

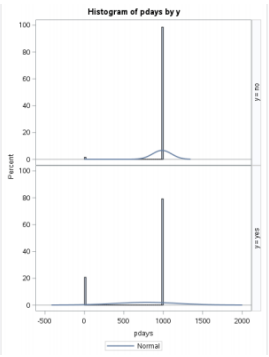
No hay missings en este dataset. Un vistazo a los gráficos revela que la edad del cliente no tiene influencia en la variable objetivo (y); los histogramas estan centrados en la misma región y tienen varianza parecida sin importar el valor de y.

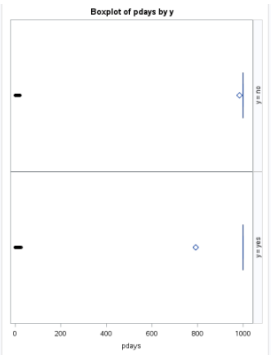
Las siguientes clasificaciones se hacen en base a los histogramas, diagramas de cjas y frecuencias de cada varible pues estos revelan que pueden ser agrupadas. Además, esto será más económico computacionalmente.

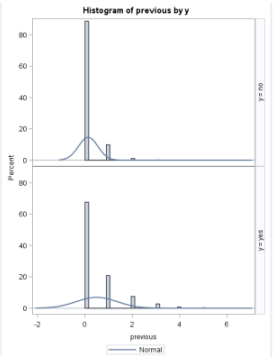
Continuous Variable	Figure	Figure Notes	Categorization Notes	Categorized Variable
campaign cons_price_idx_cat	1	> 97% de los datos bajo los10 niveles para ambos niveles de respuesta.	Ordinal en {1,2,3,>3)	campaign_cat
pdays	2	> 79% del dato pdays = 999 (no contacto previo) Para los dos niveles de respuesta.	Binario en {'contacted before' y 'never contacted'}	pdays_cat
previous	3	>68% del dato en los nive-les mas pequeños para Ambos niveles.	Binarios en {'contacted before' y 'never contacted'}	previous_cat
emp_var_rate	4	Por los graficos de barras, La clasificación parece{<=-1.8, (-1.8 to -0.1], > -0.1} para ambos niveles.	Ordinal en {<=-1.8, (-1.8 to -0.1], > -0.1}	emp_var_rate_cat
cons_price_idx	5	Datos multimodales agrupados en cuartiles.	'cons.price.idx'n < 93.056333333 — 8992 — 0.21831601 93.056333333 <= 'cons.price.idx'n < 93.911666667 — 11966 — 0.29052151 93.911666667 <= 'cons.price.idx'n — 20230 — 0.49116247	cons_price_idx_cat
cons_conf_idx	6	Datos multimodales organizados en sixtiles.	euribor3m < 1.2991788 — 8636 — 0.20967272 1.2991788 <= euribor3m < 4.1910304 — 8430 — 0.20467126 4.1910304 <= euribor3m < 4.864149 — 8446 — 0.20505973 4.864149 <= euribor3m < 4.9620732 — 8498 — 0.20632223 4.9620732 <= euribor3m — 7178 — 0.17427406	cons_conf_idx_cat

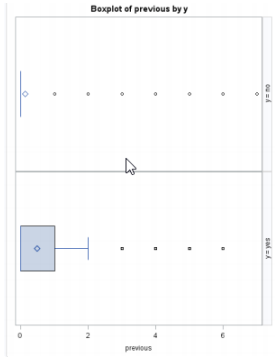
euribor3m	7	Datos multimodales organizados en sixtiles.	euribor3m < 1.2991788 | 8636 | 0.20967272 1.2991788 <= euribor3m < 4.1910304 | 8430 | 0.20467126 4.1910304 <= euribor3m < 4.864149 | 8446 | 0.20505973 4.864149 <= euribor3m < 4.9620732 | 8498 | 0.20632223 4.9620732 <= euribor3m | 7178 | 0.17427406	euribor3m_cat
nr_employed	8	Datos multimodales organizados en cuartiles.	'nr.employed'n < 5099.10335 | 13498 | 0.32771681 5099.10335 <= 'nr.employed'n < 5191.0171 | 7773 | 0.18872002 5191.0171 <= 'nr.employed'n | 19917 | 0.48356317	nr_employed_cat

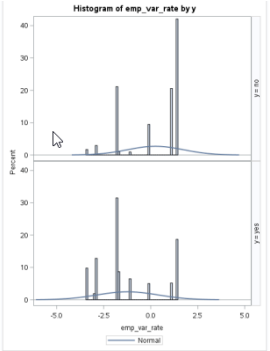


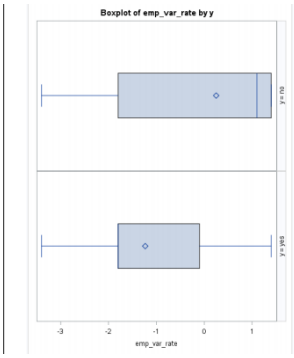


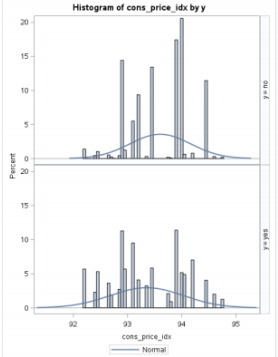


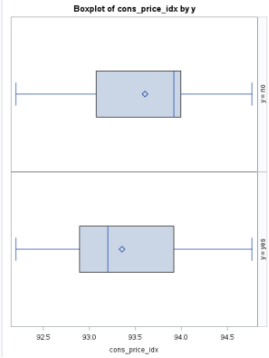


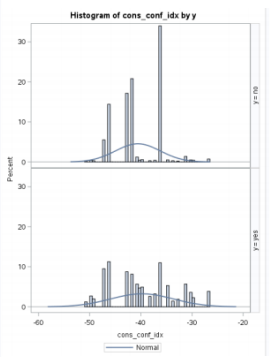


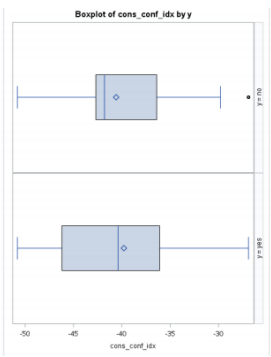


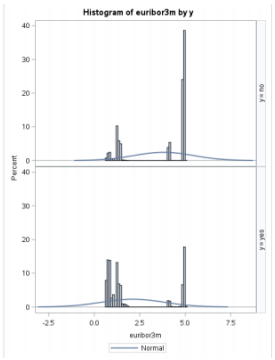


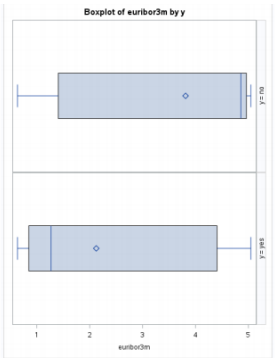


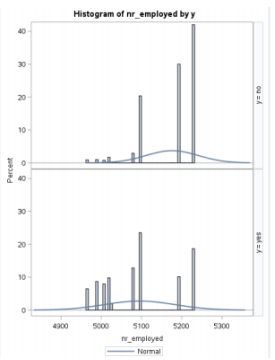


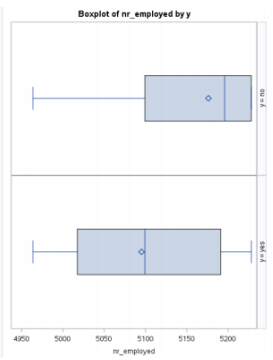












4.2. categóricas

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, campaign_cat, previous_cat, poutcome, emp_var_rate_cat, cons_price_idx_cat,	Sin variar.

cons_conf_idx_cat, euribor3m, nr_employed_cat	
day_of_week	Sin variar.

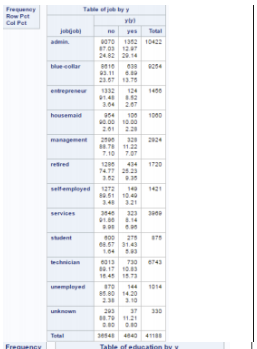


Table of job by y

job(job)	y(y) no	y(y) yes	Total
admin.	9070 87.03 24.82	1352 12.97 29.14	10422
blue-collar	8616 93.11 23.57	638 6.89 13.75	9254
entrepreneur	1332 91.48 3.64	124 8.52 2.67	1456
housemaid	954 90.00 2.61	106 10.00 2.28	1060
management	2596 88.78 7.10	328 11.22 7.07	2924
retired	1286 74.77 3.52	434 25.23 9.35	1720
self-employed	1272 89.51 3.48	149 10.49 3.21	1421
services	3646 91.86 9.98	323 8.14 6.96	3969
student	600 68.57 1.64	275 31.43 5.93	875
technician	6013 89.17 16.45	730 10.83 15.73	6743
unemployed	870 85.80 2.38	144 14.20 3.10	1014
unknown	293 88.79 0.80	37 11.21 0.80	330
Total	36548	4640	41188

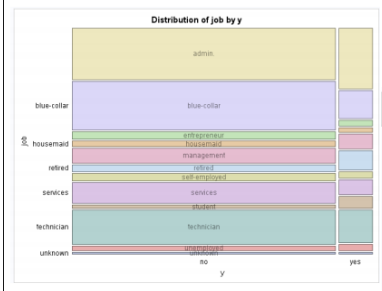
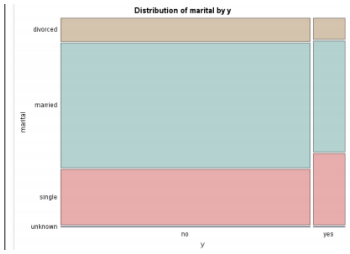


Table of marital by y

marital(marital)	y(y) no	y(y) yes	Total
divorced	4136 89.68 11.32	476 10.32 10.26	4612
married	22396 89.84 61.28	2532 10.16 54.57	24928
single	9948 86.00 27.22	1620 14.00 34.91	11568
unknown	68 85.00 0.19	12 15.00 0.26	80
Total	36548	4640	41188



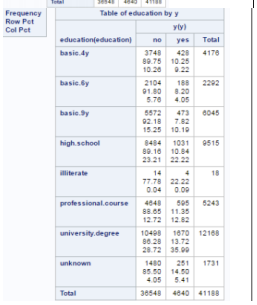


Table of education by y

education(education)	y(y) no	y(y) yes	Total
basic.4y	3748 89.75 10.26	428 10.25 9.22	4176
basic.6y	2104 91.80 5.76	188 8.20 4.05	2292
basic.9y	5572 92.18 15.25	473 7.82 10.19	6045
high.school	8484 89.16 23.21	1031 10.84 22.22	9515
illiterate	14 77.78 0.04	4 22.22 0.09	18
professional.course	4648 88.65 12.72	595 11.35 12.82	5243
university.degree	10498 86.28 28.72	1670 13.72 35.99	12168
unknown	1480 85.50 4.05	251 14.50 5.41	1731
Total	36548	4640	41188

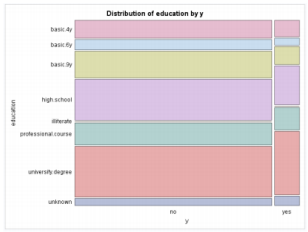
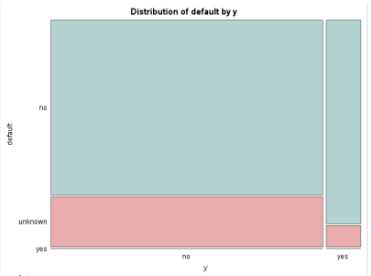


Table of default by y

default(default)	y(y) no	y(y) yes	Total
no	28391 87.12 77.68	4197 12.88 90.45	32588
unknown	8154 94.85 22.31	443 5.15 9.55	8597
yes	3 100.00 0.01	0 0.00 0.00	3
Total	36548	4640	41188



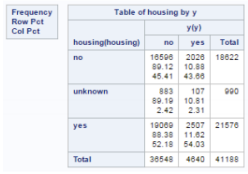


Table of housing by y

housing(housing)	y(y) no	y(y) yes	Total
no	16596 89.12 45.41	2026 10.88 43.66	18622
unknown	883 89.19 2.42	107 10.81 2.31	990
yes	19069 88.38 52.18	2507 11.62 54.03	21576
Total	36548	4640	41188

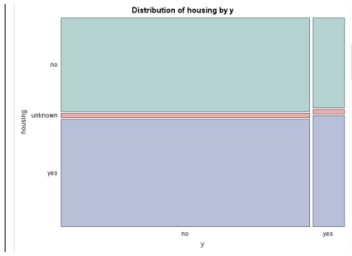
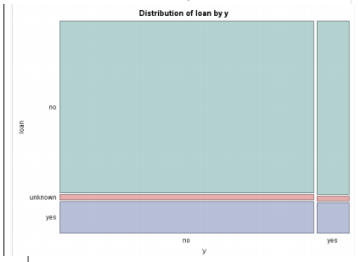


Table of loan by y

loan(loan)	y(y) no	y(y) yes	Total
no	30100 88.66 82.36	3850 11.34 82.97	33950
unknown	883 89.19 2.42	107 10.81 2.31	990
yes	5565 89.07 15.23	683 10.93 14.72	6248
Total	36548	4640	41188



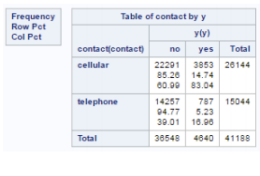


Table of contact by y

contact(contact)	y(y) no	y(y) yes	Total
cellular	22291 85.26 60.99	3853 14.74 83.04	26144
telephone	14257 94.77 39.01	787 5.23 16.96	15044
Total	36548	4640	41188

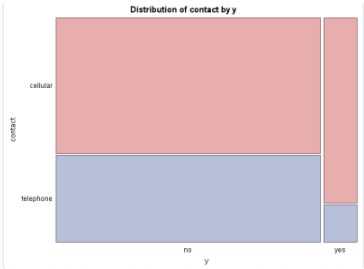
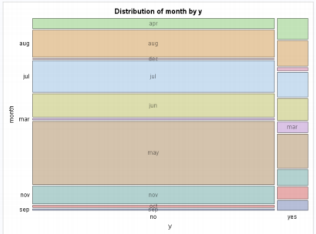


Table of month by y

month(month)	y(y) no	y(y) yes	Total
apr	2093 79.52 5.73	539 20.48 11.62	2632
aug	5523 89.40 15.11	655 10.60 14.12	6178
dec	93 51.10 0.25	89 48.90 1.92	182
jul	6525 90.95 17.85	649 9.05 13.99	7174
jun	4759 89.49 13.02	559 10.51 12.05	5318
mar	270 49.45 0.74	276 50.55 5.95	546
may	12883 93.57 35.25	886 6.43 19.09	13769
nov	3685 89.86 10.08	416 10.14 8.97	4101
oct	403 56.13 1.10	315 43.87 6.79	718
sep	314 55.09 0.86	256 44.91 5.52	570
Total	36548	4640	41188



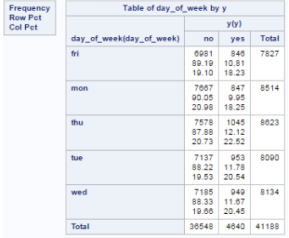


Table of day_of_week by y

day_of_week(day_of_week)	y(y) no	y(y) yes	Total
fri	6981 89.19 19.10	846 10.81 18.23	7827
mon	7667 90.05 20.98	847 9.95 18.25	8514
thu	7578 87.88 20.73	1045 12.12 22.52	8623
tue	7137 88.22 19.53	953 11.78 20.54	8090
wed	7185 88.33 19.66	949 11.67 20.45	8134
Total	36548	4640	41188

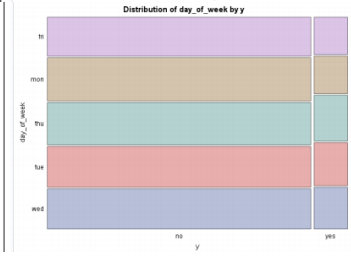
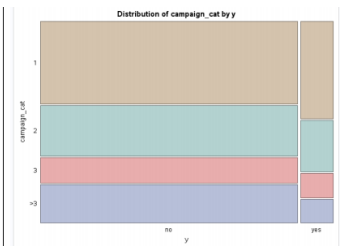
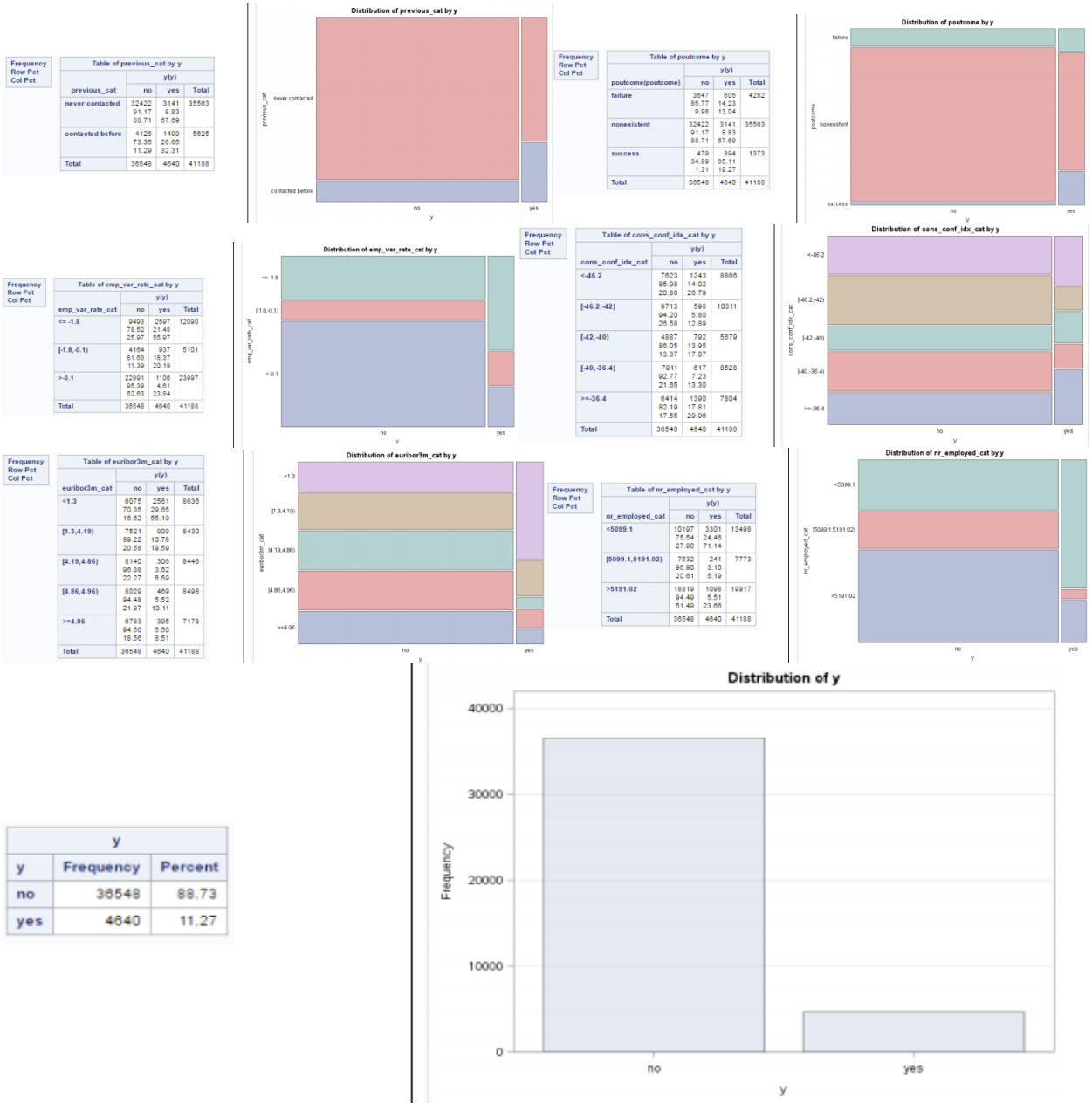


Table of campaign_cat by y

campaign_cat	y(y) no	y(y) yes	Total
1	15342 86.96 41.98	2300 13.04 49.57	17642
2	9359 88.54 25.61	1211 11.46 26.10	10570
3	4767 89.25 13.04	574 10.75 12.37	5341
>3	7080 92.73 19.37	555 7.27 11.96	7635
Total	36548	4640	41188





Frequency / Row Pct / Col Pct	Table of previous_cat by y			
		y(y)		
previous_cat		no	yes	Total
never contacted		32422 91.17 88.71	3141 8.83 67.69	35563
contacted before		4126 73.36 11.29	1499 26.65 32.31	5625
Total		36548	4640	41188

Frequency / Row Pct / Col Pct	Table of poutcome by y			
		y(y)		
poutcome(poutcome)		no	yes	Total
failure		3647 85.77 9.98	605 14.23 13.04	4252
nonexistent		32422 91.17 88.71	3141 8.83 67.69	35563
success		479 34.89 1.31	894 65.11 19.27	1373
Total		36548	4640	41188

Frequency / Row Pct / Col Pct	Table of emp_var_rate_cat by y			
		y(y)		
emp_var_rate_cat		no	yes	Total
<= -1.8		9493 78.62 25.97	2597 21.48 55.97	12090
[-1.8,-0.1)		4164 81.63 11.39	937 18.37 20.19	5101
>-0.1		22891 95.39 62.63	1106 4.61 23.84	23997
Total		36548	4640	41188

Frequency / Row Pct / Col Pct	Table of cons_conf_idx_cat by y			
		y(y)		
cons_conf_idx_cat		no	yes	Total
<-46.2		7623 85.98 20.86	1243 14.02 26.79	8866
[-46.2,-42)		9713 94.20 26.58	598 5.80 12.89	10311
[-42,-40)		4887 86.05 13.37	792 13.95 17.07	5679
[-40,-36.4)		7911 92.77 21.65	617 7.23 13.30	8528
>=-36.4		6414 82.19 17.55	1390 17.81 29.96	7804
Total		36548	4640	41188

Frequency / Row Pct / Col Pct	Table of euribor3m_cat by y			
		y(y)		
euribor3m_cat		no	yes	Total
<1.3		6075 70.36 16.62	2561 29.65 55.19	8636
[1.3,4.19)		7521 89.22 20.58	909 10.78 19.59	8430
[4.19,4.86)		8140 96.38 22.27	306 3.62 6.59	8446
[4.86,4.96)		8029 94.48 21.97	469 5.52 10.11	8498
>=4.96		6783 94.50 18.56	395 5.50 8.51	7178
Total		36548	4640	41188

Frequency / Row Pct / Col Pct	Table of nr_employed_cat by y			
		y(y)		
nr_employed_cat		no	yes	Total
<5099.1		10197 75.54 27.90	3301 24.46 71.14	13498
[5099.1,5191.02)		7532 96.90 20.61	241 3.10 5.19	7773
>5191.02		18819 94.49 51.49	1098 5.51 23.66	19917
Total		36548	4640	41188

y		
y	Frequency	Percent
no	36548	88.73
yes	4640	11.27

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:

Procedimiento FREQ

Frecuencia	Tabla de Effect por ProbChiSq		
	ProbChiSq(Pr > Chi-cuadrado)		
Effect(Efecto)	<.0001	1.0000	Total
cons.price.idx	15	5	20
default	10	0	10
job	15	0	15
loan	5	0	5
Total	45	5	50

Efecto	RCuadrado
loan	0.377712
cons.price.idx	0.318861
default	0.318861
job	0.312566
cons.price.idx	0.302027
default	0.286958
job	0.267239
cons.price.idx	0.242964
job	0.217528
cons.price.idx	0.174834

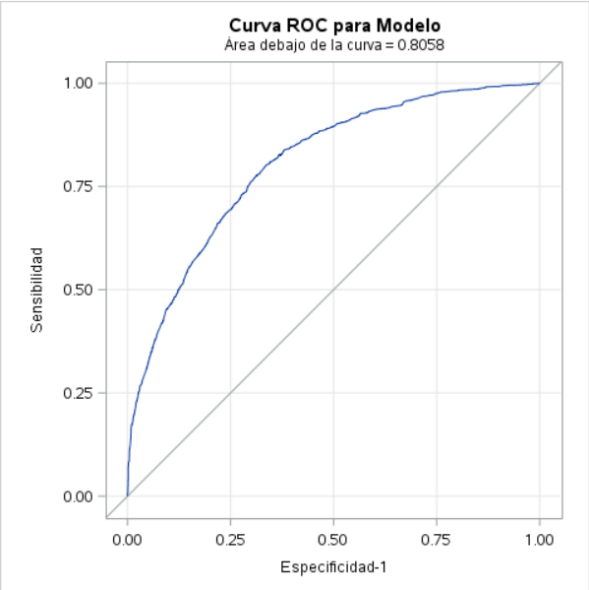
Efecto	Error estándar
cons.price.idx	0.0293
job	0.0679
cons.price.idx	0.1322
job	0.1782
cons.price.idx	0.1829
default	0.2232
loan	0.2430
default	0.2766
job	0.4116
cons.price.idx	1.0547

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.



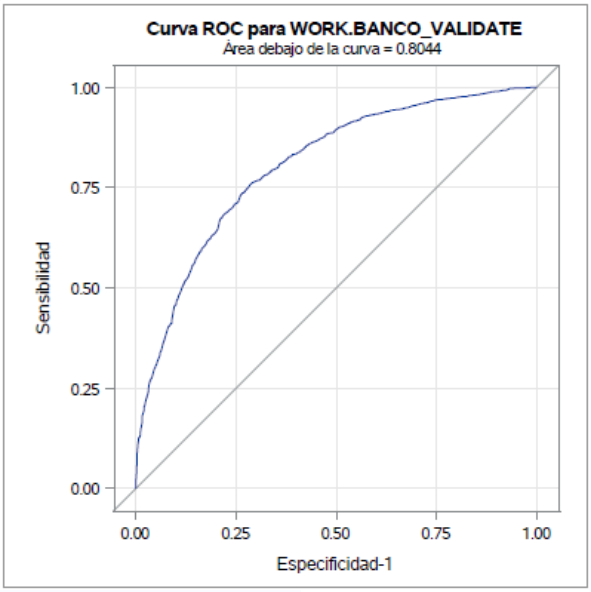


Tabla de clasificación									
	Correcto		Incorrecto		Porcentajes				
Nivel de prob	Evento	No-evento	Evento	No-evento	Correcto	Sensi-bilidad	Especi-ficidad	Falso POS	Falso NEG
0.050	2295	56	2295	6	50.5	99.7	2.4	50.0	9.7
0.100	2279	298	2053	22	55.4	99.0	12.7	47.4	6.9
0.150	2254	532	1819	47	59.9	98.0	22.6	44.7	8.1
0.200	2204	731	1620	97	63.1	95.8	31.1	42.4	11.7
0.250	2136	972	1379	165	66.8	92.8	41.3	39.2	14.5
0.300	2076	1126	1225	225	68.8	90.2	47.9	37.1	16.7
0.350	1990	1328	1023	311	71.3	86.5	56.5	34.0	19.0
0.400	1903	1468	883	398	72.5	82.7	62.4	31.7	21.3
0.450	1785	1595	756	516	72.7	77.6	67.8	29.8	24.4
0.500	1684	1685	666	617	72.4	73.2	71.7	28.3	26.8
0.550	1529	1809	542	772	71.8	66.4	76.9	26.2	29.9
0.600	1351	1907	444	950	70.0	58.7	81.1	24.7	33.3
0.650	1182	2029	322	1119	69.0	51.4	86.3	21.4	35.5
0.700	1055	2100	251	1246	67.8	45.8	89.3	19.2	37.2
0.750	869	2177	174	1432	65.5	37.8	92.6	16.7	39.7
0.800	624	2258	93	1677	62.0	27.1	96.0	13.0	42.6
0.850	488	2303	48	1813	60.0	21.2	98.0	9.0	44.0
0.900	267	2331	20	2034	55.8	11.6	99.1	7.0	46.6
0.950	22	2350	1	2279	51.0	1.0	100.0	4.3	49.2
1.000	0	2351	0	2301	50.5	0.0	100.0	.	49.5

```
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;
```

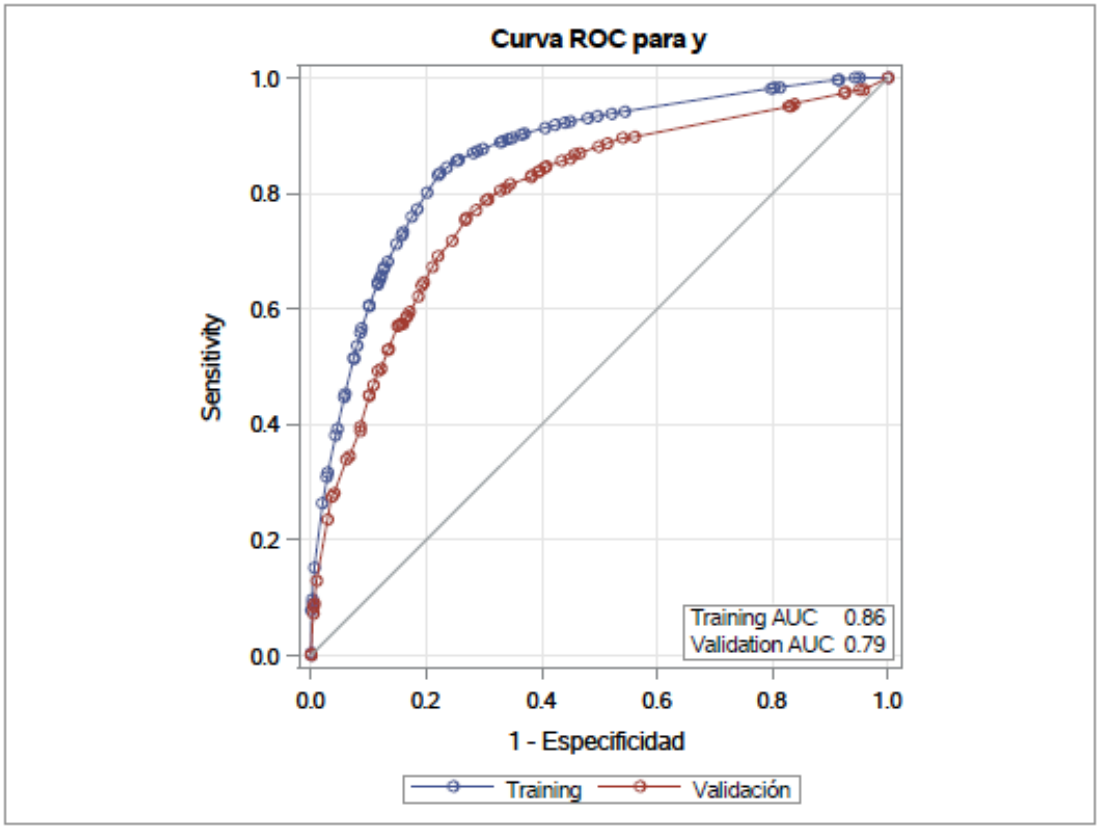
A pesar de tener un Rcuadrado mediocre, el clasificador compuesto por estas variables parece ser decente con un AUC de 0.8 tanto para training como para validación por lo quje decido quedarme con ese conjunto de caracteristicas para el modelo de regresión logística.

6.3. Otros modelos y comparación con artículo de sas

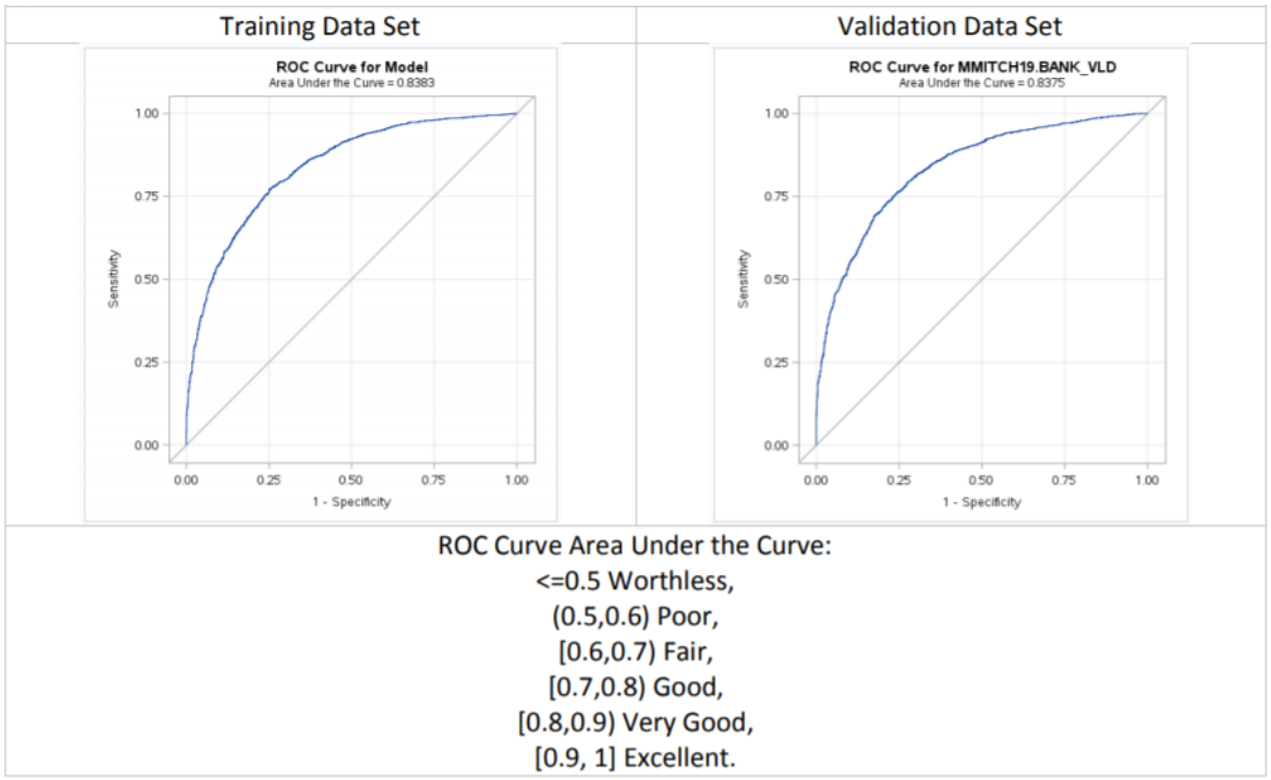
En primer lugar, he seguido el preprocesamiento que hacen porque me parece el mejor que se podría hacer a este dataset en concreto. Además los tipos de gráficos que sugieren tambien me lo parecen.

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:

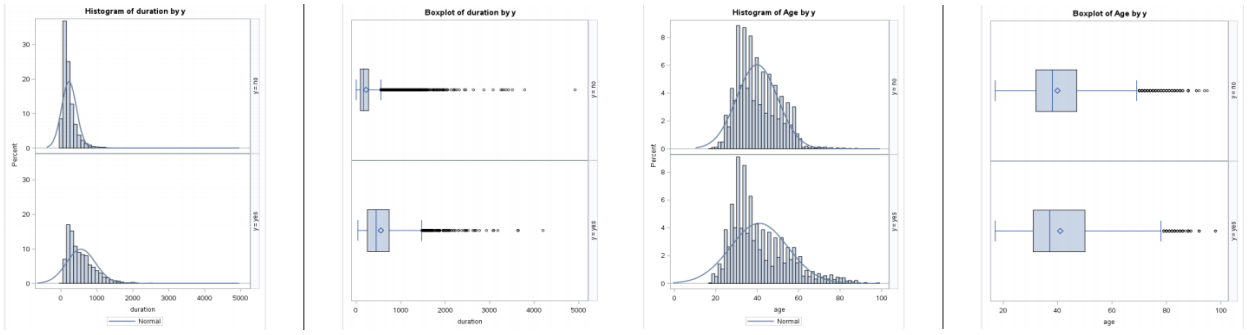


(*)



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;
```