Practica KeepCoding SAS

La práctica la he realizado con SAS estudio ya que SAS miner me daba muchos problemas.

Cocinado de los datos

En primer lugar analizaremos los datos categóricos y los numericos y los procesaremos para dejar el dataset preparado para los modelos.

Transformar variables categóricas

job :
Estudio de frecuencia:
Todos los valores tienen una cantidad de apariciones suficiente como para quedarse en el modelo, por lo que no agruparemos ninguna y las codificamos con one hot encoding

job	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
admin.	10422	25.30	10422	25.30
blue-collar	9254	22.47	19676	47.77
entrepreneur	1456	3.54	21132	51.31
housemaid	1060	2.57	22192	53.88
management	2924	7.10	25116	60.98
retired	1720	4.18	26836	65.15
self-employed	1421	3.45	28257	68.60
services	3969	9.64	32226	78.24
student	875	2.12	33101	80.37
technician	6743	16.37	39844	96.74
unemployed	1014	2.46	40858	99.20
unknown	330	0.80	41188	100.00

marital:
Estudio de frecuencia

Mismo caso que el anterior, codificamos con one hot encoding

marital	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
divorced	4612	11.20	4612	11.20
married	24928	60.52	29540	71.72
single	11568	28.09	41108	99.81
unknown	80	0.19	41188	100.00

Education:
Estudio de frecuencia

En este caso los valores de las variables tienen un orden lógico, por lo que lo codificamos con enteros incrementales. Como los desconocidos son bastantes y no tienen un orden, le dare un valor intermedio.

education	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
basic.4y	4176	10.14	4176	10.14
basic.6y	2292	5.56	6468	15.70
basic.9y	6045	14.68	12513	30.38
high.school	9515	23.10	22028	53.48
illiterate	18	0.04	22046	53.53
professional.course	5243	12.73	27289	66.25
university.degree	12168	29.54	39457	95.80
unknown	1731	4.20	41188	100.00

Default:
Esta variable puede tener poco valor ya que la mayoría son No o desconocidos.
Aun asi lo pasamos a one hot encoding antes de descartarla

default	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
no	32588	79.12	32588	79.12
unknown	8597	20.87	41185	99.99
yes	3	0.01	41188	100.00

Housing
Lo transformo a one hot encoding para mantener los 3 valores

housing	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
no	18622	45.21	18622	45.21
unknown	990	2.40	19612	47.62
yes	21576	52.38	41188	100.00

Loan
Lo transformo a one hot encoding para mantener los 3 valores

loan	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
no	33950	82.43	33950	82.43
unknown	990	2.40	34940	84.83
yes	6248	15.17	41188	100.00

Contact
Lo transformo a one hot encoding para mantener los 3 valores

contact	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
cellular	26144	63.47	26144	63.47
telephone	15044	36.53	41188	100.00

Month
Lo transformo a numérico donde a cada me le doy su numero correspondiente

month	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
apr	2632	6.39	2632	6.39
aug	6178	15.00	8810	21.39
dec	182	0.44	8992	21.83
jul	7174	17.42	16166	39.25
jun	5318	12.91	21484	52.16
mar	546	1.33	22030	53.49
may	13769	33.43	35799	86.92
nov	4101	9.96	39900	96.87
oct	718	1.74	40618	98.62
sep	570	1.38	41188	100.00

Day_of_week
Lo transformo a numérico, donde a cada día le doy un valor del 1 al 5, ya que tienen un orden.

day_of_week	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
fri	7827	19.00	7827	19.00
mon	8514	20.67	16341	39.67
thu	8623	20.94	24964	60.61
tue	8090	19.64	33054	80.25
wed	8134	19.75	41188	100.00

Analisis de las variables numéricas

Missing values

Con el procedimiento FREQ hemos observado que no hay missing values en ninguna variable

Tratamiento de outliers

Con el procedimiento means, sacamos los estadísticos de las variables numéricas

Procedimiento MEANS

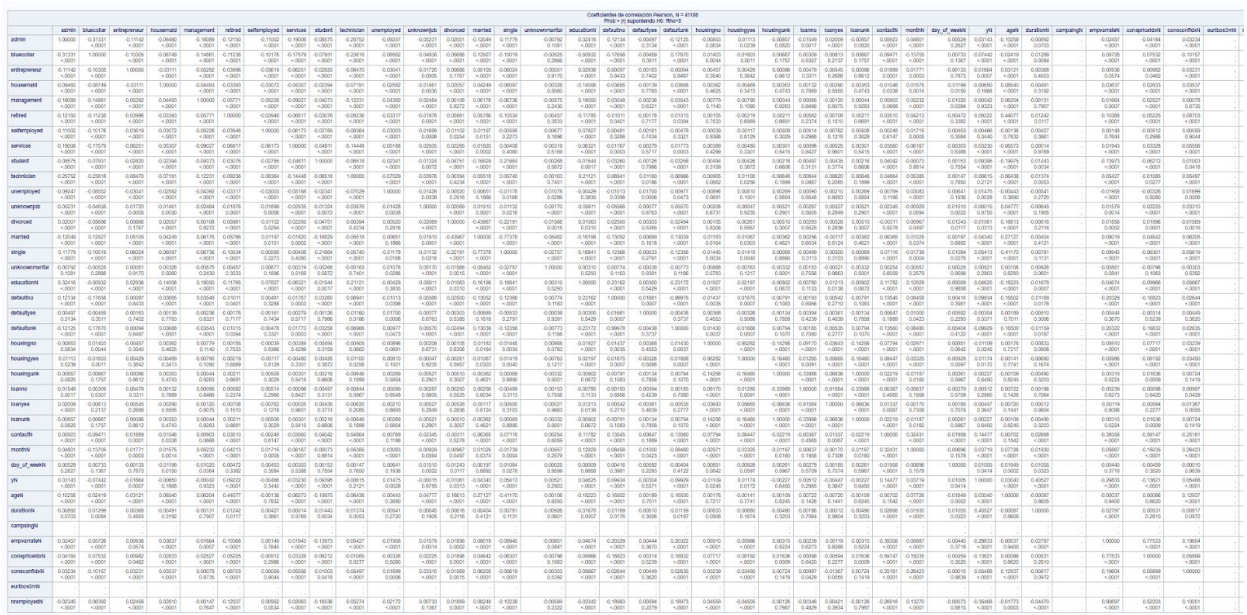
Variable	Media	Cuartil inferior	Cuartil superior	N	Máximo	Mínimo
age	40.0240604	32.0000000	47.0000000	41188	98.0000000	17.0000000
duration	258.2850102	102.0000000	319.0000000	41188	4918.00	0
campaign	2.5675925	1.0000000	3.0000000	41188	56.0000000	1.0000000
pdays	962.4754540	999.0000000	999.0000000	41188	999.0000000	0
previous	0.1729630	0	0	41188	7.0000000	0
emp.var.rate	0.0818855	-1.8000000	1.4000000	41188	1.4000000	-3.4000000
cons.price.idx	93.5756644	93.0750000	93.9940000	41188	94.7670000	92.2010000
cons.conf.idx	-40.5026003	-42.7000000	-36.4000000	41188	-26.9000000	-50.8000000
euribor3m	3.6212908	1.3440000	4.9610000	41188	5.0450000	0.6340000
nr.employed	5167.04	5099.10	5228.10	41188	5228.10	4963.60

Todo lo que supere Q3 + 1.5RIC o sea menor Q1 - 1.5RIC lo sustituiremos por esos respectivos valores para eliminar los outliers (los valores los he calculado por fuera de SAS porque no he sabido hacerlo en SAS)
En previous y pdays no tiene sentido hacer este tratamiento

	Q3 + 1.5RIC	Q1 - 1.5RIC
age	80,75	-1,75
duration	644,5	-223,5
campaing	6	-2
pdays	999	999
previous	0	0
emp.var.rate	6,2	-6,6
cons.price.idx	95,374	91,695
cons.conf.idx	-26,95	-52,15
euribos3m	10,391	-4,086
nr.employed	5421,6	4905,6

Correlacion

Analizamos la tabla de correlacion de las variables



Existen algunas variables que podríamos eliminar por tener alta correlación pero de momento las dejamos

Análisis de las variables desde el punto de vista del negocio

Desde el punto de vista del negocio parece que las siguientes variables pueden tener más peso:

- Edad: La edad parece que puede ser un factor importante para pedir un depósito
- Que no tengas prestamos previos (loan, housing)
- La estabilidad de su trabajo
- La duración del último contacto nos puede hacer pensar si ha podido tener algún interés

Estas serán las variables candidatas de nuestros datos procesados para nuestros modelos:

- entrepreneur management student selfemployed housingno loanno ageN durationN

Modelos

Modelo de regresión lineal.

Utilizaré la macro de validación cruzada para evaluar varios modelos y me quedaré con el que menor error tenga. Aparte de las variables previamente elegidas se incorporaran otras en los diferentes modelos para evaluar la mejora.

Procedimiento MEANS

Variable	N	Media	Desv. est.	Mínimo	Máximo
suma	20	3399.14	0.5947224	3398.00	3400.21
semilla	20	1243.50	5.9160798	1234.00	1253.00
modelo	20	1.0000000	0	1.0000000	1.0000000

Procedimiento MEANS

Variable	N	Media	Desv. est.	Mínimo	Máximo
suma	20	3355.83	0.7357942	3354.43	3357.86
semilla	20	1243.50	5.9160798	1234.00	1253.00
modelo	20	2.0000000	0	2.0000000	2.0000000

Procedimiento MEANS

Variable	N	Media	Desv. est.	Mínimo	Máximo
suma	20	4019.44	0.7244305	4018.14	4021.04
semilla	20	1243.50	5.9160798	1234.00	1253.00
modelo	20	3.0000000	0	3.0000000	3.0000000

De los 3 modelos elegidos, parece que el que mejor se comporta es el segundo, que tiene las variables previamente seleccionadas y alguna más relacionadas sobre todo con el trabajo: entrepreneur management student selfemployed housingno loanno ageN durationN educationN services retired housemaid

Una vez seleccionado nuestro modelo óptimo pasamos a evaluarlo

Procedimiento REG
Modelo: MODEL1
Variable dependiente: yN

N.º observaciones leídas	41188
N.º observaciones usadas	41188

Análisis de varianza

Origen	DF	Suma de cuadrados	Cuadrado de la media	Valor F	Pr > F
Modelo	12	764.52595	63.71050	782.42	<.0001
Error	41175	3352.75870	0.08143		
Total corregido	41187	4117.28465			

Raíz MSE	0.28535	R-cuadrado	0.1857
Media dependiente	0.11265	R-Sq Ajust	0.1854
Coef Var	253.30110		

Podemos observar que tiene una bondad de ajuste de 0.18 que no es un buen valor y la suma de los errores tiene un valor alto. Usaremos estos parámetros para comparar con otros modelos

Modelo GLM

Para el modelo GLM he utilizado el procedimiento glmselect, he introducido las variables que he comentado en el modelo anterior y he introducido interacciones con laedad. El método de seleccion de variables ha sido stepwise.

El resultado ha sido el siguiente:

Procedimiento GLMSELECT
Modelo seleccionado

El modelo seleccionado es el modelo en el último paso (Paso 13).

Efectos:	Intercept entrepreneur student selfemployed durationN educationN retired housemaid student*ageN selfemployed*ageN housingno*ageN ageN*retired ageN*services ageN*housemaid

El mejor modelo lo ha encontrado en el paso 13 y estos son los efectos que ha considerado

 entrepreneur student selfemployed durationN educationN retired housemaid student*ageN selfemployed*ageN housingno*ageN ageN*retired ageN*services ageN*housemaid

Análisis de varianza				
Origen	DF	Suma de cuadrados	Cuadrado de la media	Valor F
Modelo	13	813.39950	62.56919	779.76
Error	41174	3303.88515	0.08024	
Total corregido	41187	4117.28465		

Raíz MSE	0.28327
Media dependiente	0.11265
R-cuadrado	0.1976
R-Sq Ajust	0.1973
AIC	-62701
AICC	-62701
BIC	-103889
C(p)	15.02846
PRESS	3307.39799
SBC	-103771
ASE	0.08021

La bondad de ajuste del modelo es de 0.19, o que tampoco es muy bueno.

Escojo los efectos seleccionados y lo aplico a un modelo GLM

Parámetro	Estimación	Error estándar	t valor	Pr > |t|
T. independiente	-.0892663132	0.00521612	-17.11	<.0001
entrepreneur	-.0254211460	0.00759797	-3.35	0.0008
student	0.6018649153	0.05069269	11.87	<.0001
selfemployed	0.0733163606	0.03283908	2.23	0.0256
durationN	0.0004926478	0.00000539	91.48	<.0001
educationN	0.0138201528	0.00087584	15.78	<.0001
retired	-.7455049755	0.04099807	-18.18	<.0001
housemaid	-.2815716381	0.03777517	-7.45	<.0001
student*ageN	-.0155935605	0.00192077	-8.12	<.0001
selfemployed*ageN	-.0021389375	0.00079876	-2.68	0.0074
ageN*housingno	-.0001998843	0.00006674	-3.00	0.0027
retired*ageN	0.0145143607	0.00065175	22.27	<.0001
ageN*services	-.0005235916	0.00012257	-4.27	<.0001
housemaid*ageN	0.0066110248	0.00080829	8.18	<.0001

Ninguna de las variables tiene un p-valor superior a 0.05 por lo que no elimino ninguna variable

Modelo redes neuronales

Utilizaré la macro para efectuar validación cruzada y comparar varios modelos

Las variables iniciales son las mismas a las seleccionadas en otros modelo. Cambiamos parámetros como la función de activación y el número de nodos para encontrar la mejor configuración del modelo.

El tiempo de procesamiento es muy alto, por lo que no he podido realizar muchas prebas para optimizar el modelo

Procedimiento MEANS

Analysis Variable : suma						
modelo	N Obs	N	Media	Desv. est.	Mínimo	Máximo
Modelo 1	1	1	3357.62	.	3357.62	3357.62
Modelo 2	1	1	3177.26	.	3177.26	3177.26

El modelo 2 tener la suma de los errores menor que el modelo 1, por lo que será el escogido

Aquí podemos observar los resultados de la optimización del modelo

Resultados de optimización			
Iteraciones	10	Invocaciones de función	27
Invocaciones del gradiente	12	Restricciones activas	0
Función objetivo	0.2554746165	Elemento gradiente absoluto máx	0.0001579628
Coeficiente angular de la dirección de búsqueda	-8.648172E-8	Radio	1

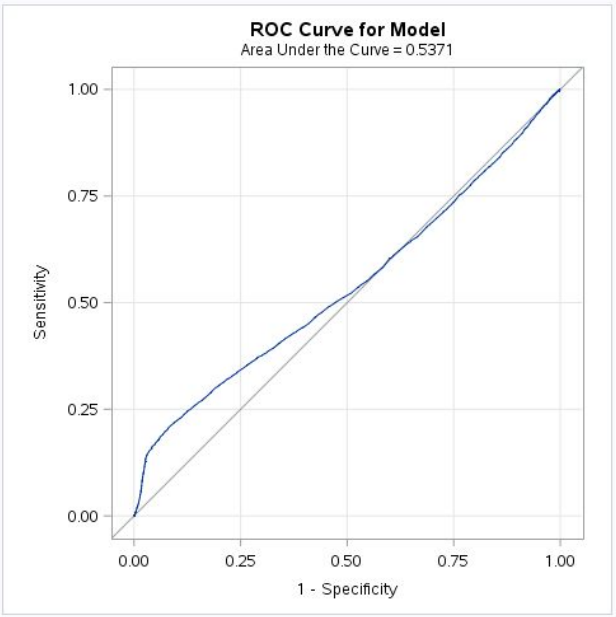
Regresión logística

A continuación vamos a probar con la macro de regresión logistica. E tenido que reducir el numero de variables porque tardaba demasiado en ejecutarse.

El mejor modelo incorpora las siguientes variables

entrepreneur management student selfemployed housingno loanno ageN

La curva ROC de este modelo nos muestra que es un modelo muy malo



Adjunto también los estadísticos de calidad del modelo.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	29000.724	28631.664
SC	29009.350	28700.671
-2 Log L	28998.724	28615.664

Conclusiones

Para hacer una buena elección de modelo habría que haber dedicado más tiempo a la elección de las variables y haber dejado más tiempo de computación, ya que todos los modelos conseguidos han sido bastante malos.

Voy a elegir el modelo GLM como modelo final porque la suma de los errores es ligeramente menor que en los otros casos.

Predicciones

Con el modelo escogido, extraigo sus predicciones y las ordeno. Ya que estamos buscando las que estén más cerca de 1 y más lejos de 0. Posteriormente obtengo el 10% de esas observaciones (4118 observaciones) (proc sql). El resto, el 5% de observaciones las obtengo de manera aleatoria.(PROC SURVEYSELECT)