Practica SAS : Modelo de venta cruzada en banco
Autor: Rafael Lucena Martinez

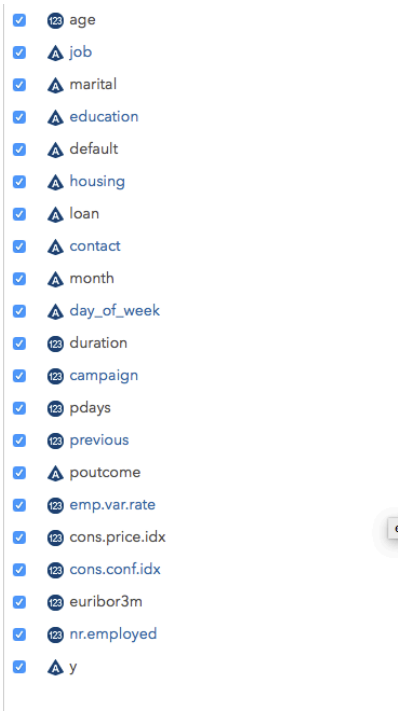
Análisis univariable de los principales estadísticos

Preparar una herramienta para análisis de gráficos exploratorios

El análisis univariable y el análisis exploratorio lo vamos a juntar en un análisis completo de cada una de las variables tanto estadístico como exploratorio.

Para poder realizar dicho análisis de la forma más óptima posible es necesario realizar un estudio previo del dataset según la naturaleza de cada una de las variables.

Nos encontramos un dataset con 20 variables independientes y una variable (y) dependiente del resto. Cada variable tiene 41188 observaciones.



Lo primero que vamos a ver en la descripción del dataset BANK_ADDITIONAL_FULL es la existencia de variables que no serán útiles. Tras estudiar el dataset comprobamos que existe una variable no predictiva (*duration*). Es decir, no podemos conocer la duración de la llamada antes de ponernos en contacto con un determinado cliente por lo que no nos servirá de nada introducir dicha variable, ya que es una variable que conoceremos a futuro. Existe otra variable con un problema similar que es *campaign*, que podemos resolver restándole uno y de esta forma ya la podríamos utilizar en nuestros modelos predictivos.

Por lo que tras quitar la variable *duration* nos encontramos con un dataset con 19 variables independientes.

De estas variables se tiene que verificar si existe o no algún missing, para lo que utilizamos *proc means* observando que no existe ninguno.

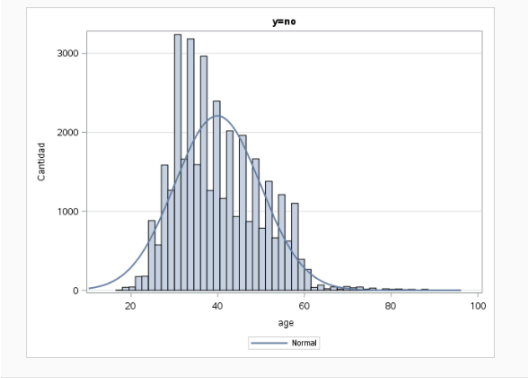
Procedimiento MEANS

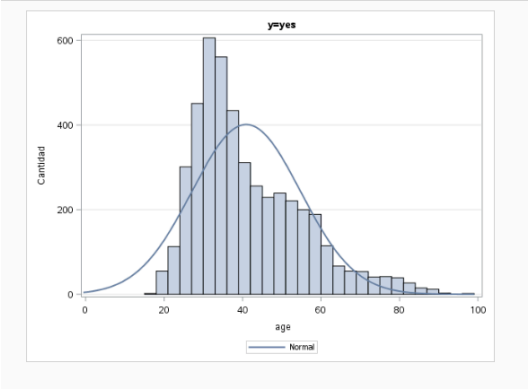
Variable	N	N Miss
age	41188	0
duration	41188	0
campaign	41188	0
pdays	41188	0
previous	41188	0
emp.var.rate	41188	0
cons.price.idx	41188	0
cons.conf.idx	41188	0
euribor3m	41188	0
nr.employed	41188	0

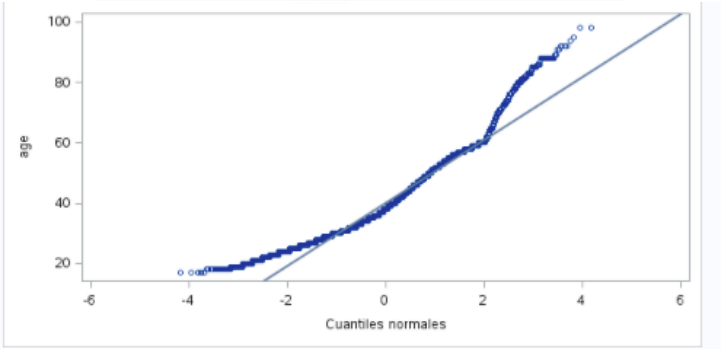
Realizamos un análisis exploratorio de cada una de las variables para ver si tenemos que algún tipo más de modificación a realizar sobre el dataset. Este estudio se realizará mediante la opción de dibujar gráficos que dispone SAS Studio por cuestión de rapidez computacional y consistirá en realizar los histogramas de las variables continuas y los diagramas de barras de las variables discretas agrupados con respecto a los dos valores de nuestra variable dependiente (yes/no), lo que nos proporcionará información muy importante de la forma y calidad de cada una de las variables lo que nos permitirá tomar decisiones su posterior tratamiento.
Para complementar el análisis exploratorio, también se ha utilizado el procedimiento *univariate* para ver los principales estadísticos que tienen cada una de las variables, algo necesario para la detección de outliers y otras anomalías.

Análisis exploratorio de variables continuas

- Age

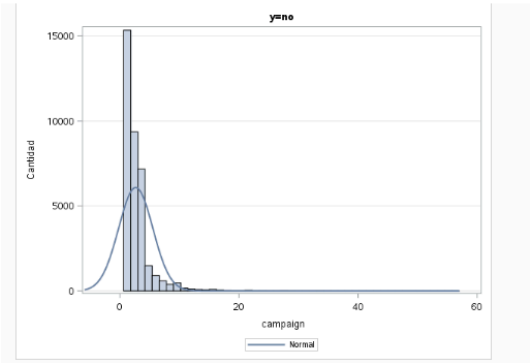


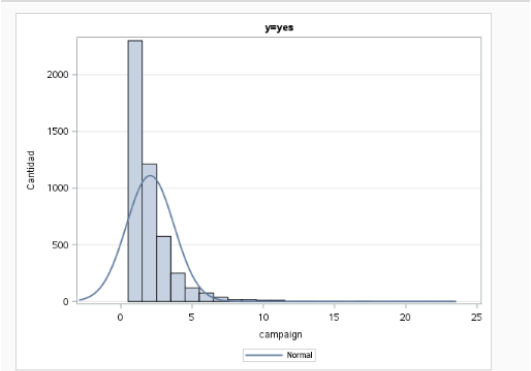


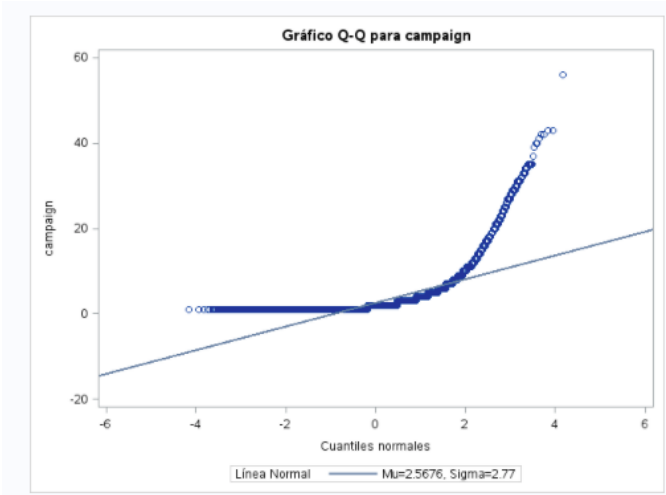


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	40.02406	Desviación std	10.42125
Mediana	38.00000	Varianza	108.60245
Moda	31.00000	Rango	81.00000
		Rango intercuartil	15.00000

- campaign

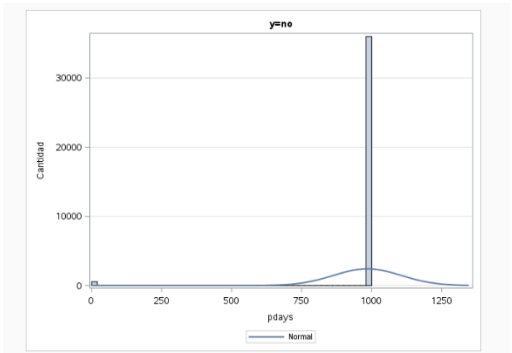


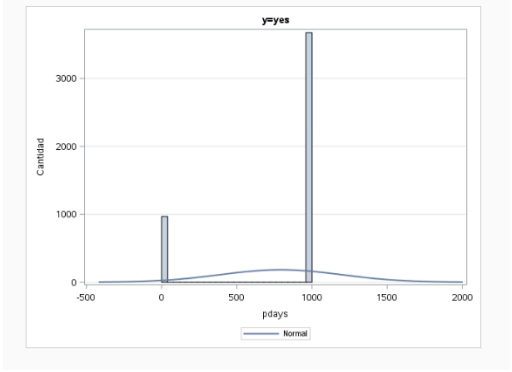


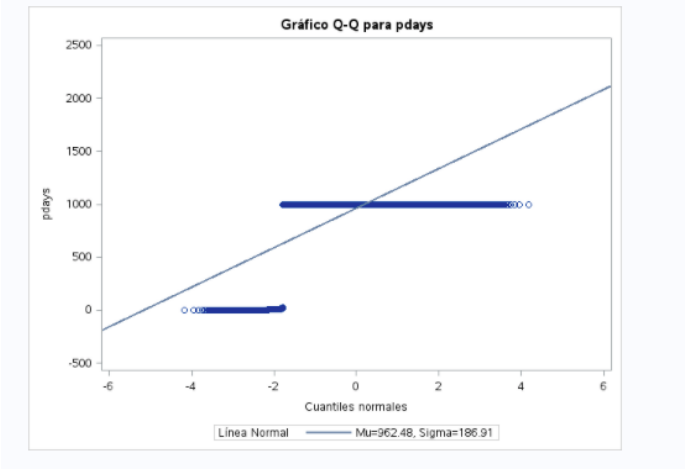


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	2.567593	Desviación std	2.77001
Mediana	2.000000	Varianza	7.67298
Moda	1.000000	Rango	55.00000
		Rango intercuartil	2.00000

- pdays

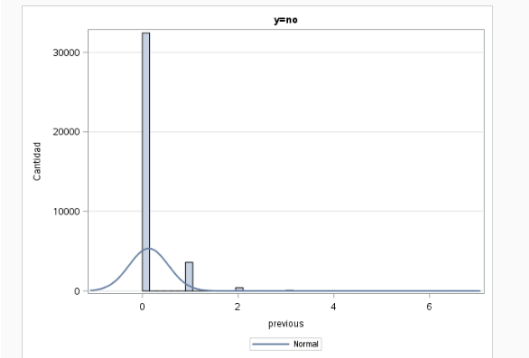


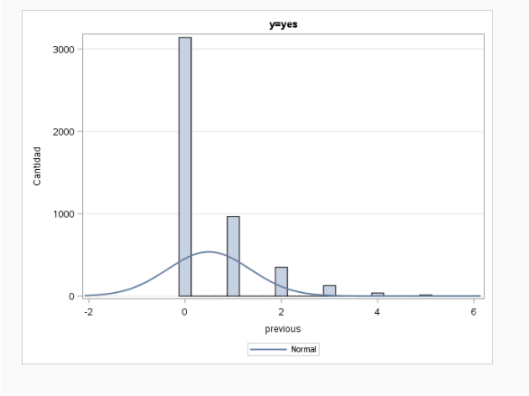


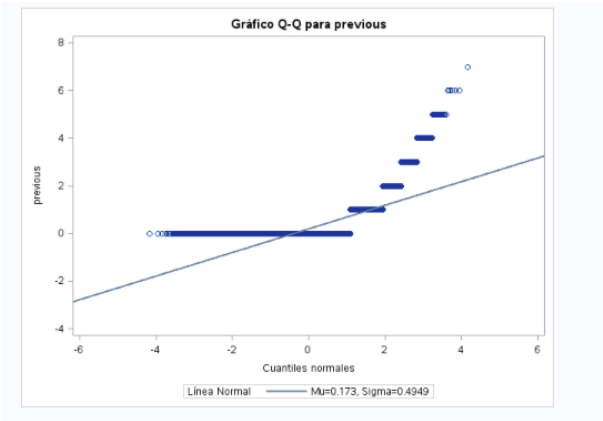


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	962.4755	Desviación std	186.91091
Mediana	999.0000	Varianza	34936
Moda	999.0000	Rango	999.00000
		Rango intercuartil	0

- previous

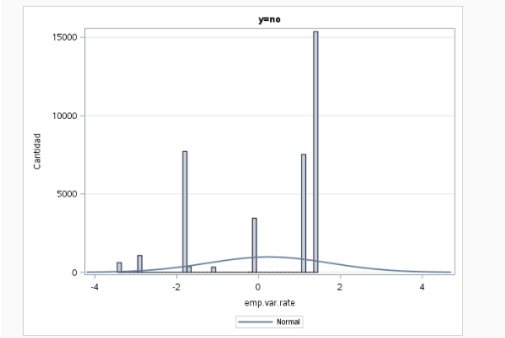


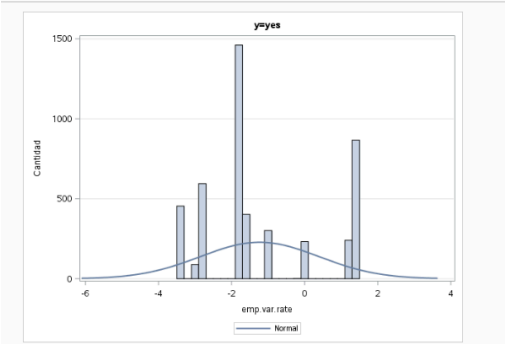


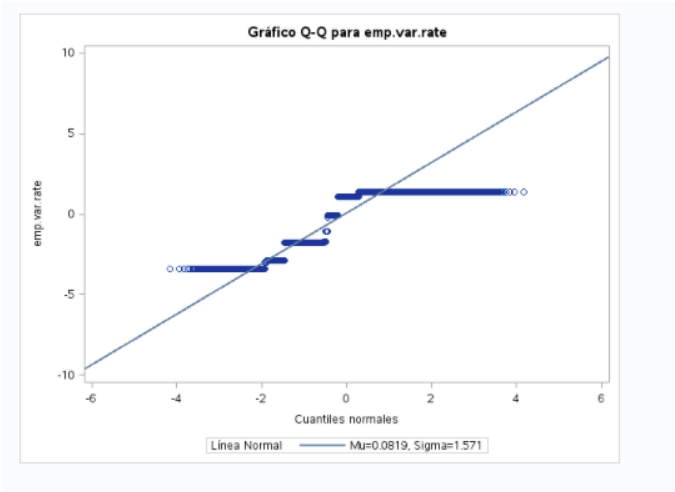


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	0.172963	Desviación std	0.49490
Mediana	0.000000	Varianza	0.24493
Moda	0.000000	Rango	7.00000
		Rango intercuartil	0

- emp.var.rate

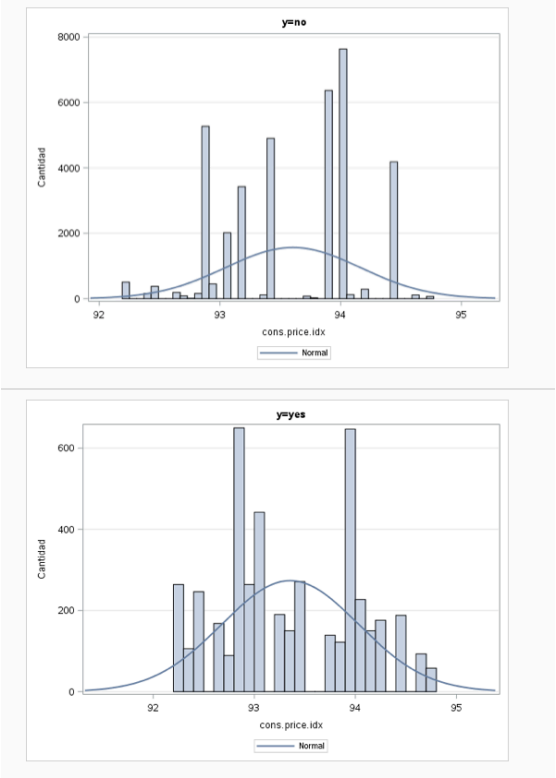


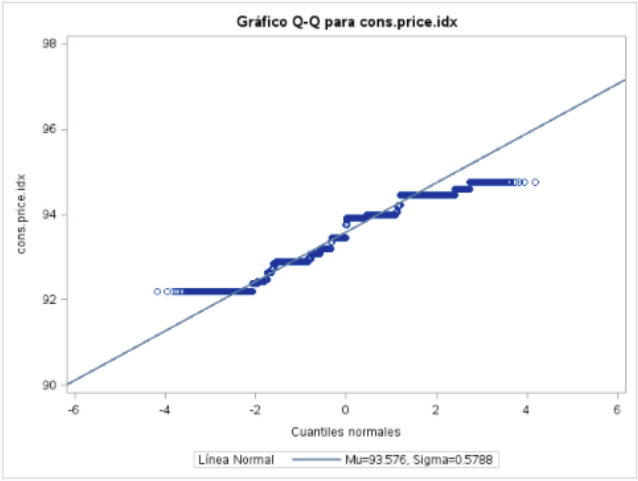




Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	0.081886	Desviación std	1.57096
Mediana	1.100000	Varianza	2.46791
Moda	1.400000	Rango	4.80000
		Rango intercuartil	3.20000

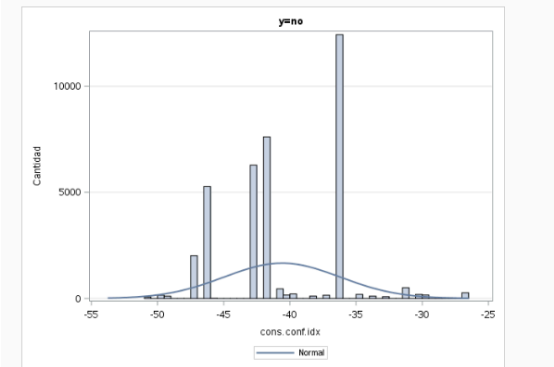
- cons.price.idx

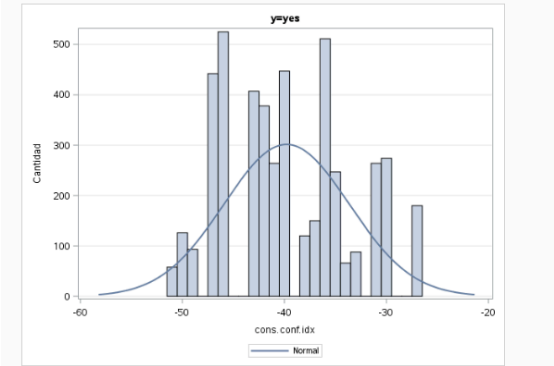


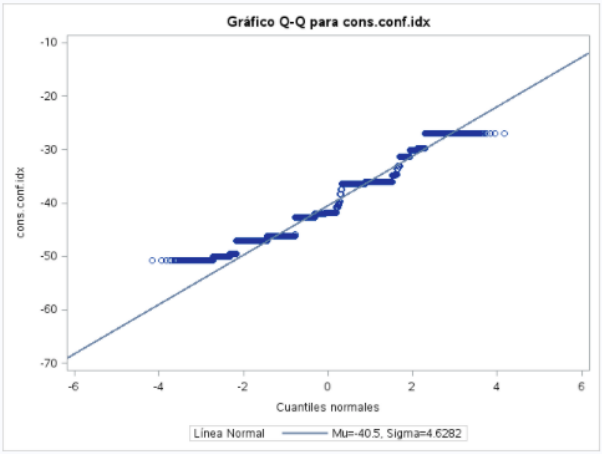


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	93.57566	Desviación std	0.57884
Mediana	93.74900	Varianza	0.33506
Moda	93.99400	Rango	2.56600
		Rango intercuartil	0.91900

- cons.conf.idx

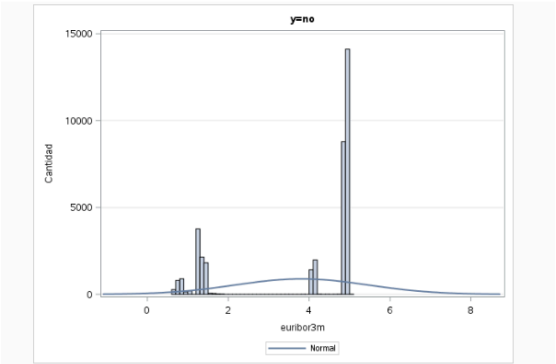


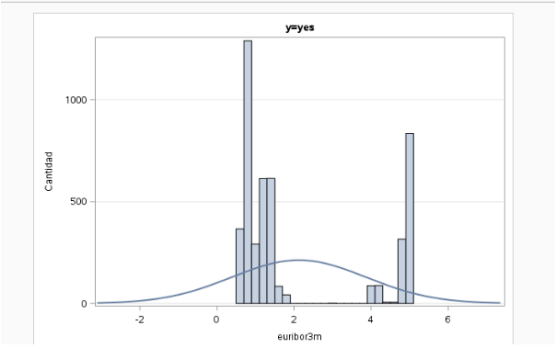


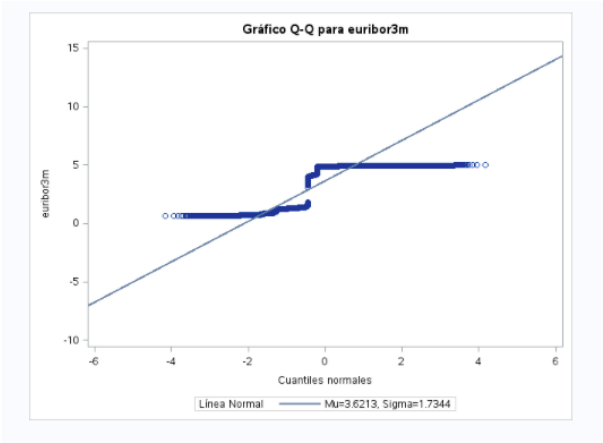


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	-40.5026	Desviación std	4.62820
Mediana	-41.8000	Varianza	21.42022
Moda	-36.4000	Rango	23.90000
		Rango intercuartil	6.30000

- euribor3m

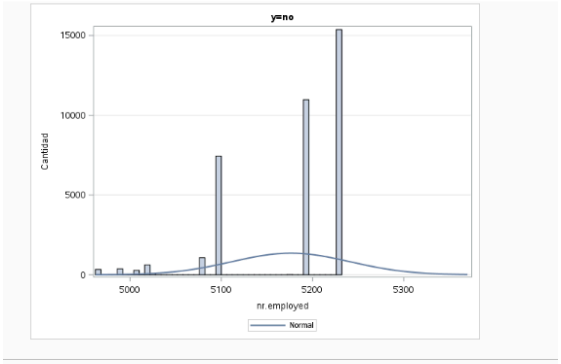


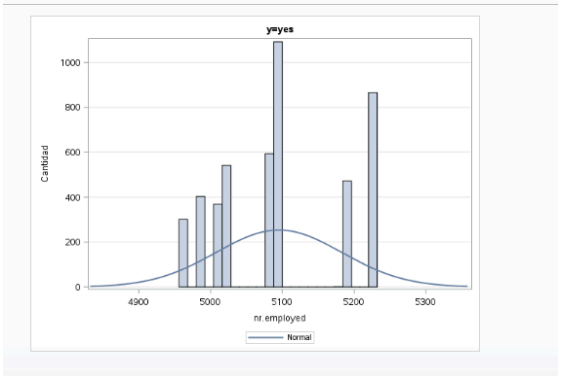


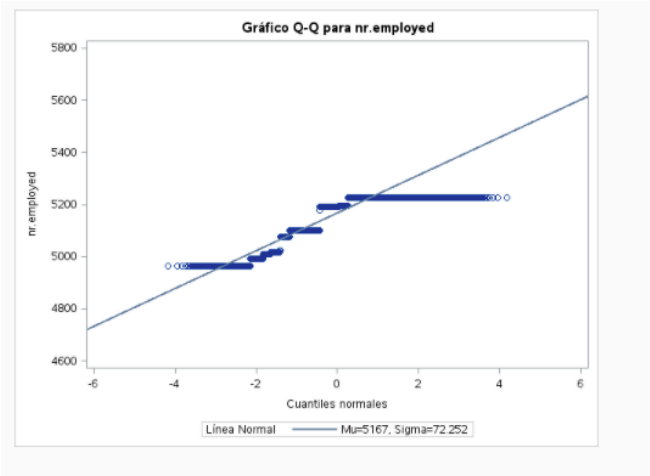


Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	3.621291	Desviación std	1.73445
Mediana	4.857000	Varianza	3.00831
Moda	4.857000	Rango	4.41100
		Rango intercuartil	3.61700

- nr.employed



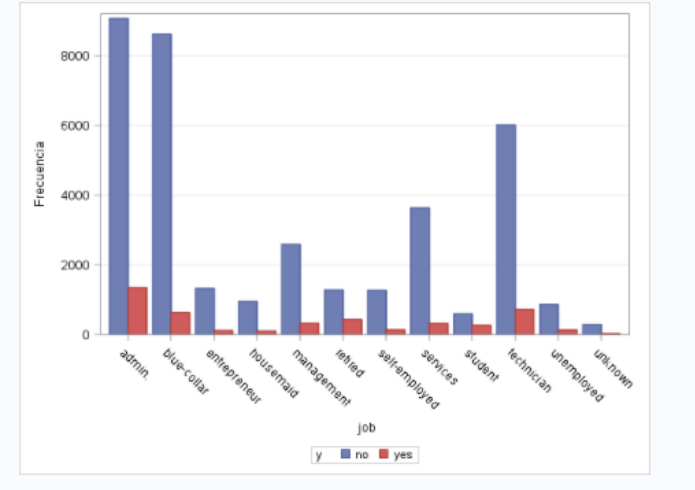




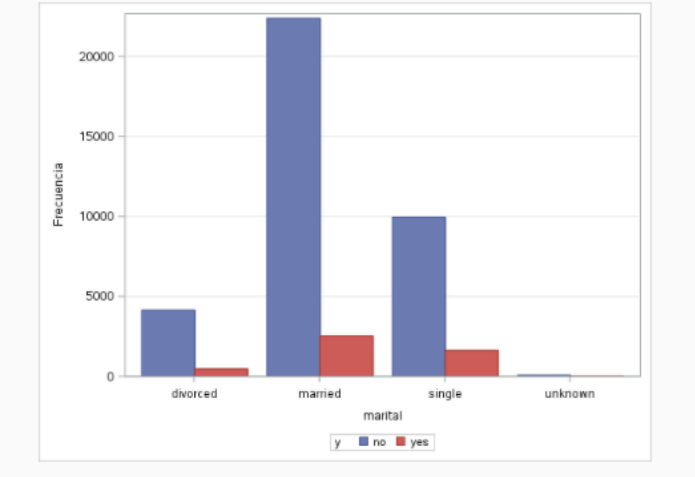
Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	5167.036	Desviación std	72.25153
Mediana	5191.000	Varianza	5220
Moda	5228.100	Rango	264.50000
		Rango intercuartil	129.00000

A continuación, obtendremos gráficos de barras de las variables discretas:

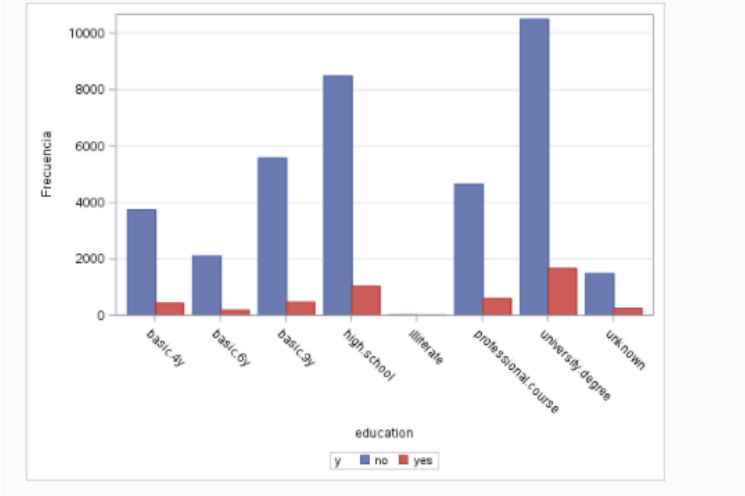
- job



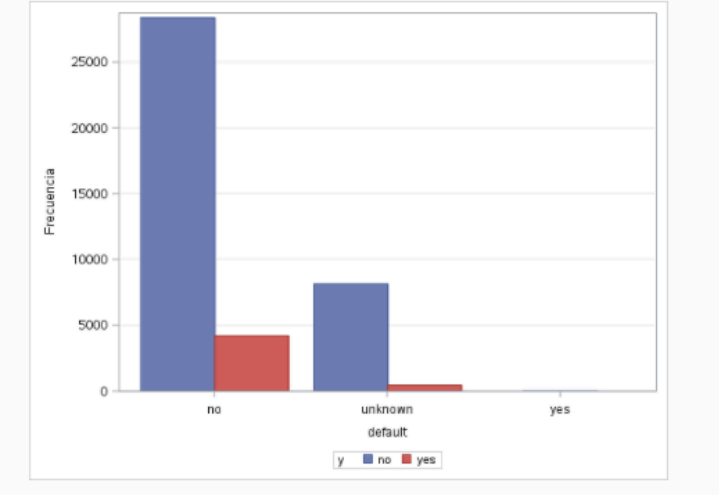
- marital



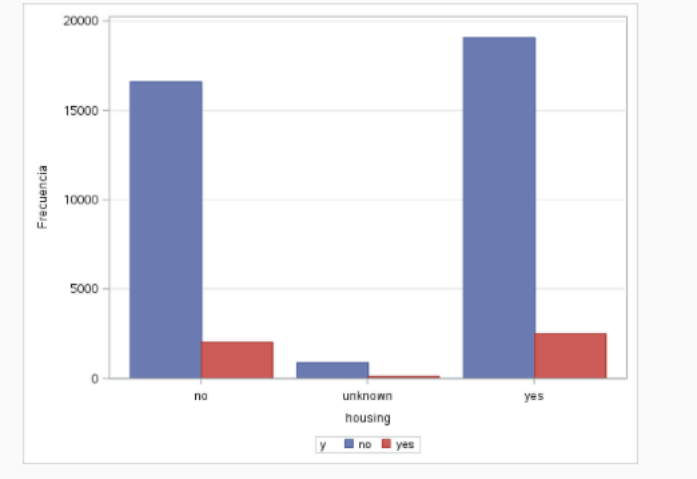
- education



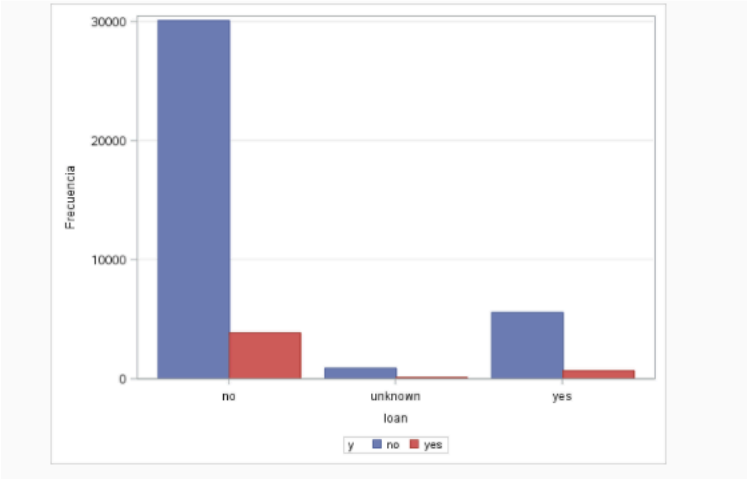
- default



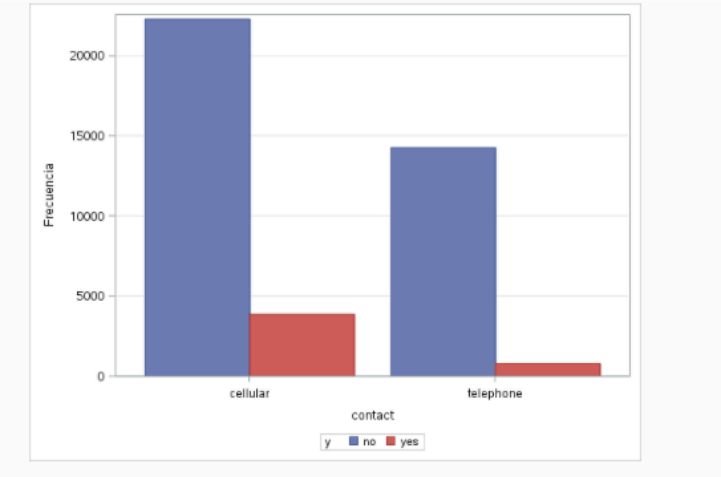
- housing



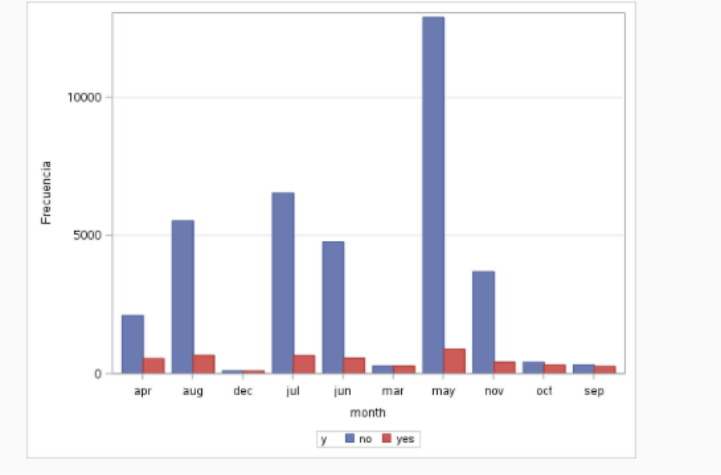
- loan



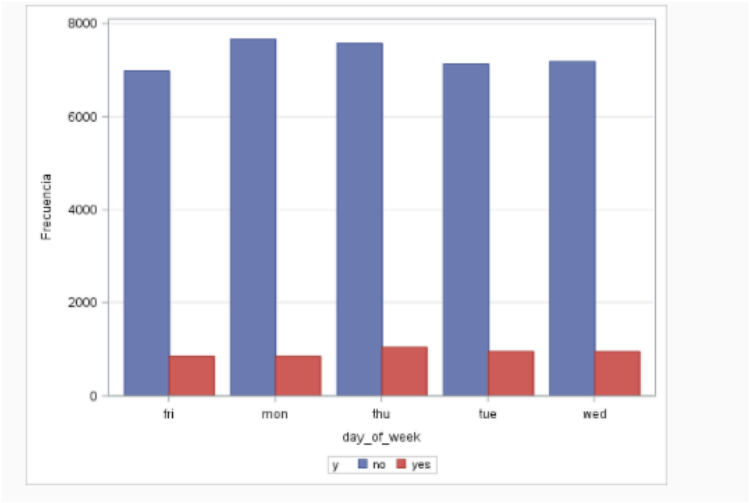
- contact



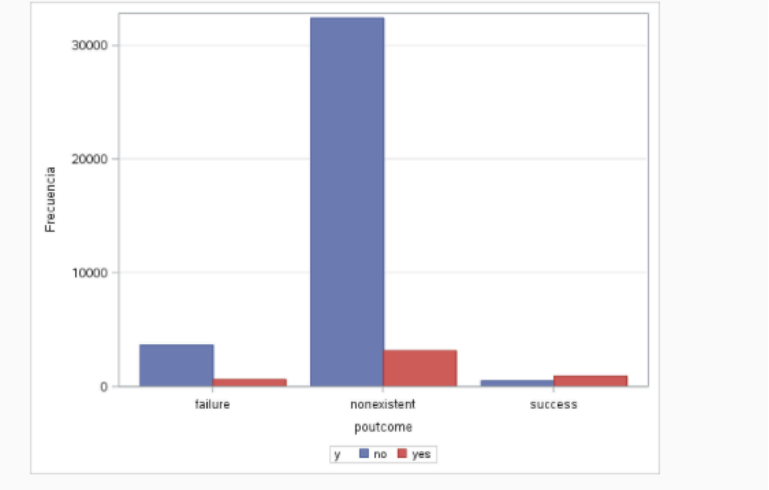
- month



- day_of_week



- poutcome



Eliminarías alguna variable del modelo

Con el análisis exploratorio y estadístico existen una serie de acciones que podemos llevar a cabo.

Existen variables que debido a la calidad del dato podemos eliminar:

-default: como se puede observar en el análisis exploratorio la calidad del dato es muy mala ya que prácticamente todos los datos son no, y no nos aportan prácticamente información. En este dataset su eliminación no tiene un impacto muy grande, pero cuando se utilizan datasets de mayor tamaño la eliminación de variables que no nos aportan información van a permitir que podamos trabajar de una forma más eficiente.

Tras eliminar esta variable nos quedaremos con un dataset con 18 variables independientes.

Existen variables que podemos tramificar sin perder prácticamente información:

-age: la podemos tramificar para de esta forma trabajar mejor con la variable
-month: la tramificamos en cuatrimestres ya que vemos que existe cuatrimestres que en los que se obtienen la mayoría de los resultados
-pdays: generamos una variable categórica, agrupando de 7 en 7 días para trabajar de forma más sencilla
-previous: hacemos binaria la variable
-campaign: restamos uno y agrupamos aquellos valores que se encuentran por encima de 4 ya que son muchos con muy pocos valores tal y como podemos comprobar en el análisis exploratorio realizado.

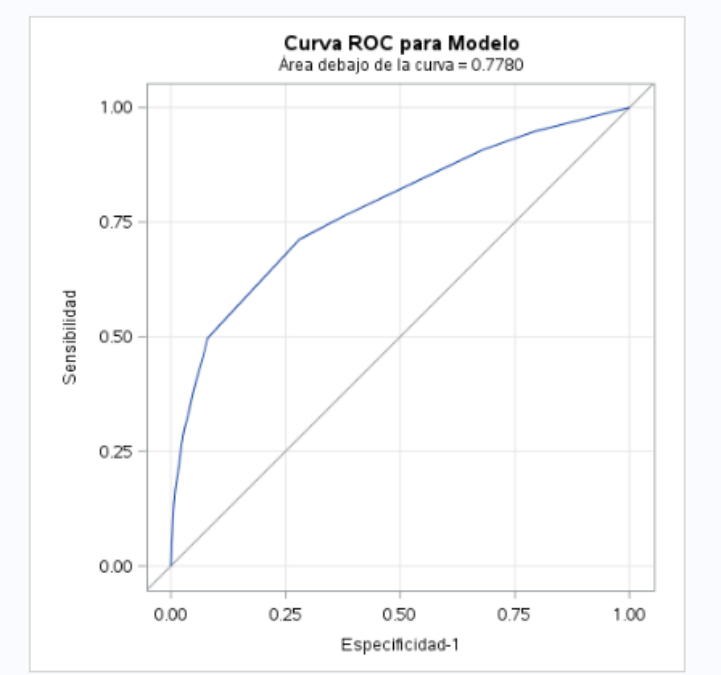
Realizar el modelo de Regresión Logística en SAS

Para realizar la regresión logística, utilizaremos la macro facilitada. Para ello es necesario convertir todas las variables de entrada en dummy. Esto se ha realizado mediante el procedimiento glmmod.

Una que se han introducido todas las variables en la macro, ésta nos selecciona las variables *emp_var_rate, pdays y quarter*. Si calculamos la tabla de sensibilidad y la curva ROC con estas variables obtenemos los siguientes resultados:

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	4211	11709	24839	429	38.7	90.8	32.0	85.5	3.5
0.100	3304	26310	10238	1336	71.9	71.2	72.0	75.6	4.8
0.150	2309	33599	2949	2331	87.2	49.8	91.9	56.1	6.5
0.200	2309	33599	2949	2331	87.2	49.8	91.9	56.1	6.5
0.250	2306	33631	2917	2334	87.3	49.7	92.0	55.8	6.5
0.300	2086	34054	2494	2554	87.7	45.0	93.2	54.5	7.0
0.350	1420	35409	1139	3220	89.4	30.6	96.9	44.5	8.3
0.400	1371	35507	1041	3269	89.5	29.5	97.2	43.2	8.4
0.450	774	36205	343	3866	89.8	16.7	99.1	30.7	9.6
0.500	774	36205	343	3866	89.8	16.7	99.1	30.7	9.6
0.550	774	36205	343	3866	89.8	16.7	99.1	30.7	9.6
0.600	774	36205	343	3866	89.8	16.7	99.1	30.7	9.6
0.650	774	36205	343	3866	89.8	16.7	99.1	30.7	9.6
0.700	641	36304	244	3999	89.7	13.8	99.3	27.6	9.9
0.750	253	36471	77	4387	89.2	5.5	99.8	23.3	10.7
0.800	0	36548	0	4640	88.7	0.0	100.0	.	11.3
0.850	0	36548	0	4640	88.7	0.0	100.0	.	11.3
0.900	0	36548	0	4640	88.7	0.0	100.0	.	11.3
0.950	0	36548	0	4640	88.7	0.0	100.0	.	11.3
1.000	0	36548	0	4640	88.7	0.0	100.0	.	11.3



Con la que obtenemos un modelo con un área debajo de la curva del 0.7780 lo cual es un valor aceptable.

Modelos de Enterprise Miner:

Regresión Lineal

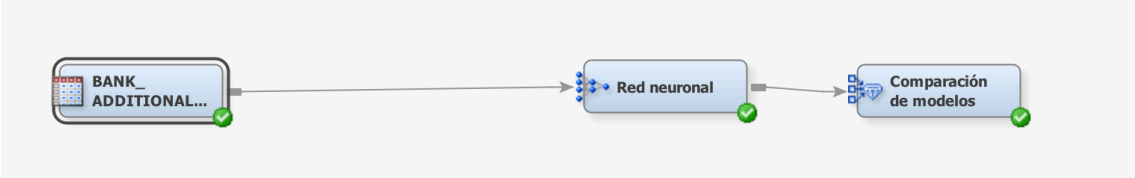
Si analizamos el problema planteado, nos encontramos con que la variable dependiente es una variable dicotómica por lo que no se puede utilizar el modelo de regresión lineal para resolver este tipo de problemas.

GLM

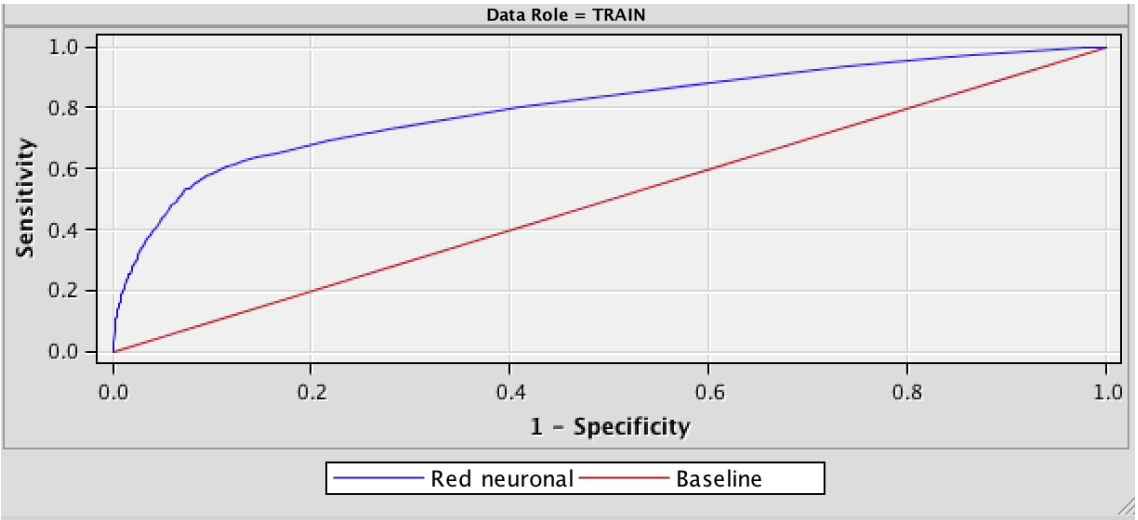
Los GLM son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc.) y varianzas no constantes. En nuestro caso la variable dependiente es una variable dicotómica, por lo que tendríamos que utilizar la distribución binomial con n=1. Esto sería aplicar una regresión logística que es lo que hemos aplicado anteriormente.

Redes Neuronales

Para realizar el modelo de redes neuronales, lo que vamos a hacer es utilizar el dataset quitando las variables duration (porque es una variable a futuro) y default (mala calidad del dato).



Al ver los resultados de este modelo, observamos los siguientes datos:



Una curva ROC con un área debajo de la misma de 0.8, lo cual es un muy buen resultado.

¿Qué modelo es mejor?

Para comparar los modelos, se ha realizado mediante la curva ROC de ambos.
De los modelos que hemos realizado, el que mejor resultados obtiene es el de redes neuronales aunque no con mucha diferencia de la regresión logística.

Realizar una selección de clientes para una campaña donde estén el 10% de los mejores clientes

Para la realización de este apartado, se ha procedido a ordenar el modelo logístico según el campo *contratado_predicted*. Una vez ordenado se cogen las 4118 observaciones que corresponde con el 10% del dataset.

Obtener el 5% de una selección aleatoria

Para obtener una muestra aleatoria se ha utilizado el procedimiento *surveyselect* que realiza esta operación de forma directa, obtiendo las 2060 observaciones aleatorias.

Para finalizar se han unido ambas tablas, recogiendo el 10% mejor y el 5% aleatorio.