Examen a casa de gerente de Ciencia de Datos

Estrategia de Clientes

Banco Azteca

Para resolver el examen por favor envía un reporte con las respuestas a las preguntas. De preferencia que el reporte sea en formato html o PDF. Puede haber sido creado en Rmarkdown, Latex, Word, Google Docs, o lo que se te facilite más. Si el documento que envías no tiene el código que usaste, por favor mándanoslo también en un archivo separado.

En el archivo *datos.csv* hay 20,000 observaciones de datos, de los cuales una muestra de 20 observaciones se ve de la siguiente forma:

y	x1	x2	x3	x4	x5	x6	x7
1	-1.3377270	-2.484148	0.3619298	1.2723203	0.2262391	R	D
0	4.3930290	-2.507107	0.6740250	-0.6064462	0.0755804	R	C
1	7.7328314	-2.390456	1.4482842	-0.3148993	0.1534020	R	D
0	4.1263868	-1.644245	1.2953372	0.0267299	0.3251762	R	C
0	2.1808083	-2.537360	0.1554671	0.5906692	0.2764094	R	C
0	4.4361409	-2.446668	0.7213905	0.3511013	0.0114505	R	C
0	2.2155835	-2.079620	0.2343344	-0.8355137	0.2036323	R	D
1	1.8210518	-1.940279	1.2631304	-2.3116282	0.2256087	AA	D
1	3.6023061	-1.826110	0.9459004	1.4037950	0.1422005	R	A
1	-5.2073200	-2.077535	0.1135520	-1.8654986	0.4811655	R	D
0	-2.4172317	-2.369142	1.1355147	-1.1758933	0.0415305	R	C
0	0.0620712	-1.878534	0.4466007	-1.4220414	0.0083352	R	B
1	-0.6494364	-2.888670	1.7966832	0.3384233	0.4015856	R	D
0	2.0475651	-2.369579	0.4659882	-0.6763717	0.2350104	R	C
1	-6.8752055	-2.304897	0.9171211	-1.3349783	0.0107622	AA	C
0	2.9716296	-2.347071	0.6616015	1.4921797	0.0803157	R	B
0	9.6436040	-1.860196	0.3645962	2.1297181	0.1459957	R	B
0	4.5991505	-2.230633	1.1983322	-0.5975851	0.0552770	R	C
1	10.9771776	-2.199348	0.5924462	0.5600820	0.5499272	R	C
0	5.0930160	-2.002046	0.4169881	2.1250246	0.0218817	R	B

Se quiere construir un modelo predictivo para la variable de interés y a partir de las variables $x1$ a $x7$. La variable y es categórica y solo toma los valores 1 o 0. Las variables $x5$ y $x6$ también son categóricas. El resto son numéricas.

1) Realiza un análisis exploratorio de datos. Puedes crear gráficas univariadas, bivariadas, resúmenes, tablas, o todo lo que creas necesario para conocer los datos que tienes a la mano.

2) Crea al menos dos modelos predictivos para la variable y a partir de las demás covariables. Trata de minimizar el número de falsos positivos y falsos negativos. Estos modelos pueden ser Compara ambos modelos utilizando lo que creas necesario como matrices de confusión, curvas ROC, gráficas, etc. ¿Qué modelo escogerías para poner en producción? ¿Por qué lo harías?

3) Si se te dijera que es preferible equivocarse en la clase 0 que en la clase 1, ¿tu decisión de la pregunta anterior sería la misma? Por favor elabora en tu respuesta.