Ejercicios para practicar: Regresión Logística

El conjunto de datos DatosVino contiene información sobre ciertas características de vinos, junto con las ventas de los mismos.

Las variables contenidas en el fichero son (observa que hay dos variables objetivo diferentes):

Variable	Descripción
Id	Código identificativo del tipo de vino
Beneficio (objetivo)	Beneficio obtenido por la venta de ese tipo de vino
Compra (objetivo)	Variable dicotómica que toma valor 1 si se ha realizado algún pedido de ese tipo de vino, y 0, en caso contrario
Acidez	Características químicas de los distintos tipos de vino: • Densidad y azúcar: sólo valores positivos. • pH: entre 4 y 10. • Restantes: rango ilimitado de valores
Acidocitrico	
Azucar	
Clorurosodico	
Densidad	
Ph	
Sulfatos	
Alcohol	Contenido de alcohol en % (debe situarse entre 0 y 100)
Etiqueta	Percepción del diseño de la etiqueta (MM=muy malo, M=malo, R=regular, B=bueno, MB=muy bueno)
CalifProductor	Calificación (entre 0 y 9) del vino según el productor.
Clasificacion	Clasificación obtenida por un equipo de expertos (4 * = excelente, 1 * = pobre)
Region	Región de la que proviene (toma 3 valores distintos)
PrecioBotella	Precio por botella

Partiendo del conjunto de datos depurado, el objetivo final de estos ejercicios es construir un modelo de regresión logística para predecir la variable Compra. Los ejercicios constan de los siguientes apartados:

1) Realiza una partición Entrenamiento-Prueba (80-20) de los datos.
2) Construye un primer modelo de regresión logística en el que incluyas todas las variables disponibles (sin las transformaciones automáticas ni las interacciones). Evalúa la calidad del modelo resultante e interpreta el parámetro de una variable continua y otra binaria.
3) Basándote en la significancia de las variables del modelo inicial, determina las variables menos útiles para predecir la variable Objetivo. A continuación, construye un modelo de regresión como el del apartado 2 pero eliminando las variables detectadas. ¿Este modelo es mejor que el anterior?
4) Basándote en la importancia de las variables del modelo inicial, determina las variables menos útiles para predecir la variable Objetivo. A continuación, construye un modelo de regresión como el del apartado 2 pero eliminando las variables detectadas. ¿Este modelo es mejor que el anterior?
5) Basándote en los resultados del V de Cramer, determina las variables menos útiles para predecir la variable objetivo. A continuación, construye un modelo de regresión como el del apartado 2 pero eliminando las variables detectadas. ¿Ha sido correcto eliminarlas? ¿Este modelo es mejor que el del apartado 3?
6) Partiendo del modelo del apartado 4, incluye las interacciones que consideres puedan ser influyentes y determina si lo son o no. ¿El modelo resultante es mejor que los anteriores?
7) Determina el mejor modelo de los anteriores según su área bajo la curva ROC.
8) Utilizando validación cruzada (20 repeticiones, 5 grupos), determina cuál de los 4 modelos anteriores es preferible basándote el área bajo la curva ROC.
9) Determina el mejor punto de corte para el modelo seleccionado en el apartado 7 según el índice de Youden. Obtén, para ese punto de corte, la tasa de acierto, sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo. Interpreta su significado.
10) Evalúa el modelo ganador (estabilidad y bondad del mismo, variables más importantes, etc.).