Ejercicios para practicar: Selección de Variables Reg. Log.

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, los ejercicios constan de los siguientes apartados:

1) Realiza una partición Entrenamiento-Prueba (80-20) de los datos.
2) Construye de nuevo el modelo ganador del segundo día de clase. Analiza los resultados y determina si existe alguna variable cualitativa cuyas categorías deban unirse. De ser así, crea una variable nueva con menos categorías y genera de nuevo el modelo pero cambiando dicha variable. ¿Observas alguna mejoría en el modelo?
3) Determina el mejor modelo de regresión lineal a partir de stepwise y backward basándote en los estadísticos AIC y el SBC incluyendo todas las variables disponibles (sin las transformaciones automáticas ni las interacciones). ¿Coinciden algunos de los 4 modelos generados? De no ser así, ¿cuál parece ser el mejor de todos?
4) Repite el ejercicio 3 (únicamente con stepwise), pero incluyendo todas las interacciones posibles.
5) Repite el ejercicio 3 (únicamente con stepwise), pero considerando las variables originales junto con las transformaciones automáticas.
6) Repite de nuevo el ejercicio 3 (únicamente con stepwise), incluyendo todas las variables, sus transformaciones automáticas y las interacciones.
7) Llegados a este punto, habrás obtenido varios modelos diferentes, es el momento de decidir cuál de todos ellos es preferible. Aplica validación cruzada repetida sobre todos ellos y determina cuál es el mejor de todos basándote en los resultados de los diagramas de cajas.
8) Lleva a cabo una selección de variables aleatoria con todas las variables (incluidas las transformaciones y las interacciones) y determina los dos modelos que más se repitan. ¿Son mejores que los modelos previamente obtenidos en el conjunto de prueba?