

## Ejercicios para practicar: Regresión Logística

El conjunto de datos VentaViviendas contiene información sobre el precio de venta de una serie de viviendas, junto con las características básicas de las mismas.

Las variables contenidas en el fichero son:

Variable	Descripción
Year, month	Año y mes de la venta
Price (objetivo)	Precio de venta de la vivienda
Luxury (objetivo)	Variable dicotómica que toma valor 1 si se trata de una vivienda de lujo (precio superior a medio millón de \$), y 0, en caso contrario.
bedrooms	Número de habitaciones
bathrooms	Número de baños (los medios se refieren a aseos)
sqft_living	Superficie del salón
sqft_lot	Superficie total (incluye jardín)
Sqft_above	Superficie excluyendo el sótano
floors	Número de plantas (los medios se refieren a entre plantas)
waterfront	¿Tiene vistas al mar? (1: si, 0: no)
view	¿Tiene buenas vistas? (1: si, 0: no)
condition	Estado de la vivienda (de A a D, siendo A el mejor estado)
basement	¿Tiene sótano? (1: si, 0: no)
yr_built	Año de construcción de la vivienda
yr_renovated	Año de renovación de la vivienda (si es 0, no ha sido renovada)
lat, long	Coordenadas de latitud y longitud de la vivienda

**Partiendo del conjunto de datos depurado, el objetivo final de estos ejercicios es construir un modelo de regresión logística para predecir la variable Luxury. 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.).