Modelos Probabilísticos y Análisis Estadístico

Taller Final Modelos de Regresión

Semestre 2020 - III

El archivo **training_set.csv** contiene la información de 17290 casas ubicadas en el condado de King County en el estado de Washington (Estados Unidos). El objetivo consiste en que usted encuentre el mejor modelo de regresión lineal múltiple que permita estimar el precio de la casa en función de las demás variables explicatorias incluidas en el archivo, excluyendo el código de la casa **id**.

Descripción de las variables:

id: Código de la casa

date: Año en que se vendió la casa

price: Precio de la casa (dólares americanos)

bedrooms: Número de habitaciones

bathrooms: Número de baños

sqft_living: Área de la casa (pies cuadrados)

sqft_lot: Área del lote (pies cuadrados)

floors: Número de pisos

waterfront: Variable dummy para determinar si la casa tenía vista a la costa o no

view: Índice de 0 a 4 de acerca de qué tan buena era la vista de la casa

condition: Condición de la casa en escala de 1 a 5, donde 1 representa una mala condición y 5 una buena condición

grade: Índice de 1 a 13, donde 1-3 indica un bajo nivel de construcción y diseño, 7 representa un nivel promedio de construcción y diseño y 11-13 indica un alto nivel de construcción y diseño

sqft_above: Área de los pisos superiores de la casa

sqft_basement: Área del sótano

yr_built: Año de construcción de la casa

yr_renovated: Año de la última renovación de la casa

zipcode: Código postal en el que se ubica la casa

lat: Latitud geográfica de ubicación de la casa

long: Longitud geográfica de ubicación de la casa

sqft_living15: Área del espacio interior de los 15 vecinos más cercanos (pies cuadrados)

sqft_loft15: Área del lote de los 15 vecinos más cercanos (pies cuadrados)

Elabore un reporte que describa todo el proceso seguido para alcanzar el que usted considere como el modelo con el mayor poder de predicción. El modelo final debe incluir la verificación de todos los supuestos que deben ser cumplidos por un modelo de regresión lineal múltiple.

El archivo **validation_set.csv** contiene el conjunto de datos para validar el modelo seleccionado por usted. Evalúe la capacidad predictiva de su modelo con este conjunto de datos a través del índice de bondad de ajuste denominado Raíz del cuadrado medio del error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}}$$