Trabajo Práctico Análisis Exploratorio y Curación de Datos

Grupo Número 15
Integrantes: Benjamin Ocampo, Matias Oria, Pamela Pairo, Antonela Sambuceti

Documentación

El presente documento describe las operaciones efectuadas en el conjunto de datos de la competencia Kaggle, sobre estimación de precios de ventas de propiedades en Melbourne, Australia. El objetivo es permitir que otros desarrolladores puedan reproducir los mismos pasos y obtener el mismo resultado.

Link al conjunto de datos a utilizar: DanB

Renombrado de columnas seleccionadas

Se renombran las columnas de la siguiente manera:

- *Bathroom* se renombra como *housing_bathroom_segment*
- *Landsize* se renombra como *housing_land_size*
- *Price* se renombra como *housing_price*
- *Rooms* se renombra como *housing_room_segment*
- *Type* se renombra como *housing_type*
- *Regionname* se renombra como *suburb_region_segment*
- *YearBuilt* se renombra como *housing_year_built*
- *BuildingArea* se renombra como *housing_bulding_area*

Exclusión de filas

1. Se eliminan aquellas filas donde la variable housing_price, asume valores alejados de su media, más allá de 2.5 veces la desviación estándar (372 filas).
2. Se elimina la fila donde la variable housing_year_built asume el valor de 1196 (1 fila).
3. Se elimina la fila donde la variable housing_bulding_area asume el valor de 44515 (1 fila).

Características seleccionadas y sus transformaciones

Categóricas

Todas las características categóricas se codifican utilizando el método Dict Vectorizer.

- *housing_bathroom_segment* (cantidad de baños): se reemplazan 34 filas de valor original 0 por el valor 1. Se agrupan sus valores más frecuentes, creando rangos que asumen 3 valores: (0,1], (1,2], (2,8].
- *housing_room_segment* (cantidad de ambientes): se agrupan sus valores más frecuentes, creando rangos que asumen 5 valores: (0,1], (1,2], (2,3], (3,4], (4,10].
- *housing_type* (tipo de vivienda): posee 3 valores posibles h (Casa), u (unidad - duplex), t: (casa adosada).

- ➢ *suburb_region_segment* (regiones): se agrupan las regiones denominadas Eastern Victoria, Northern Victoria y Western Victoria, bajo una nueva categoría denominada Victoria. Por lo tanto, la variable puede asumir 6 categorías posibles: Eastern Metropolitan, Northern Metropolitan, South-Eastern Metropolitan, Southern Metropolitan, Victoria, Western Metropolitan.

• Numéricas

Todas las características numéricas se estandarizan utilizando el método Standard Scaler.

- ➢ *housing_land_size* (tamaño del terreno)
- ➢ *housing_price* (precio de venta)
- ➢ *suburb_rental_dailyprice* (precio promedio diario de renta): se agrega el precio promedio diario de publicaciones en la plataforma AirBnB para el mismo código postal (conjunto de datos de *scrapings* del sitio realizado por Tyler Xie, disponibles en una publicación, en su perfil de Kaggle). Se seleccionaron aquellos códigos postales con una frecuencia superior a la mediana del conteo de registros. Se imputaron los valores faltantes por su valor medio.
- ➢ *housing_year_built* (año de construcción): se imputan los valores faltantes de esta columna mediante el estimador KNeighbors definido para 2 neighbors y todas las columnas previamente estandarizadas.
- ➢ *housing_bulding_area* (área de construcción): se imputan los valores faltantes de esta columna mediante el estimador KNeighbors definido para 2 neighbors y todas las columnas previamente estandarizadas.

Datos aumentados

Se agregan los 17 primeros componentes principales obtenidos a través del método de Principal Component Analysis (PCA), aplicado sobre el conjunto de datos totalmente procesado.