DOCUMENTACIÓN DE LA EXPLORACIÓN Y CURADO

Características seleccionadas

Características categóricas

1. Type: tipo de propiedad. 3 valores posibles.
2. Regionname: nombre de la región (puede pensarse como análogo a los departamentos de una provincia en Argentina). 6 valores posibles.
3. CouncilArea: área de gobierno local (puede pensarse como similar a una ciudad en Argentina). 27 valores posibles.
4. Suburb: suburbio (puede pensarse como similar a un barrio de una ciudad en Argentina). 238 valores posibles.

Todas estas categorías fueron codificadas en OneHotEncoding. Las divisiones políticas del territorio son específicas del país de origen, pero se intenta traducirlas a la realidad Argentina para interpretar mejor los datos.

Características numéricas

1. Rooms: Cantidad de habitaciones.
2. Bathroom: Cantidad de baños.
3. Car: Cantidad de cocheras.
4. Distance: Distancia al centro de la ciudad.
5. Price: precio de venta.
6. Postcode: código postal.
7. Lattitude: latitud.
8. Longitud: longitud.
9. Landsize: tamaño del terreno.
10. Propertycount: Cantidad de casas en el suburbio.
11. BuildingArea: área que ocupa la construcción.
12. Yearbuilt: año de construcción.

Las primeras 3 fueron codificadas con un método OneHotEncoding.

Criterios de exclusión de columnas

Se eliminaron las columnas:

1. 'Date','SellerG', 'Method' porque consideramos que es información no relevante para el objetivo de predecir el valor de una propiedad. Estas columnas ofrecen detalles sobre la operación (fecha, vendedor y método de venta) y no sobre la vivienda o el entorno de la misma.
2. La columna 'Bedroom2' fue analizada en un notebook de clase. Allí se observó que 'Rooms' era suficientemente representativa y que esta columna proveniente de otra fuente de datos solo agregaba ruido. Por esta razón se la excluyó.
3. En cuanto a 'Address', esta se eliminó porque se considera suficiente la información que se tiene sobre la ubicación de la vivienda. Esta columna muestra datos muy desagregados que dificultan y entorpecen

la codificación OneHot.

Criterios de exclusión de filas/ejemplo

Se excluyen valores que cumplan ciertas condiciones:

1. Rooms mayor a 5
2. Car mayor a 3
3. Distance mayor a 22
4. Yearbuilt menor a 1850
5. BuildingArea mayores a 300

Estos criterios fueron fijados en función de la distribución de los datos. Estos limites establecen que por fuera de ellos se ubican los outliers.

Transformaciones:

1. Las columnas `YearBuilt` y `BuildinArea` fueron imputadas utilizando el algoritmo `InterativeImputer` de Sklearn con un pre procesamiento tanto de escalado como de normalizado. En ambos casos se analizaron los resultados obtenidos.

Datos aumentados

1. Se agregan las 2 primeras columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado luego de estandarizar.