DIPLOMATURA EN CIENCIA DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS APLICACIONES

EXPLORACIÓN Y CURACIÓN DE DATOS

Grupo N°: 18
Integrantes:
Parada Larrosa, Francisco
Peralta, Agustín
Porcel, Carolina
Quiros, Agustina

Criterios de exclusión de ejemplos

- De la base de datos de Melbourne, se eliminan aquellos registros donde la variable precio adquiere valores atípicos (bajos y altos). Nos quedamos con valores entre $ 300.000 y $ 3.338.150 conservando un 98.9% de los registros.
- De la base de datos de AirBnB según lo pide uno de los enunciados solo se trabaja con registros que se asocien a zipcode cuya frecuencia sea mayor a 30.

Características seleccionadas

- Categóricas
    1. Suburb: Barrio donde se encuentra la propiedad.
    2. Type: Tipo de propiedad. 3 valores posibles
    3. CouncilArea: Ciudad en la que se encuentra la propiedad

    - Todas las características categóricas fueron codificadas con el método DictVectorizer creando una nueva matriz de 352 columnas y 7447 registros.
    - También se consideró como categórica la variable Postcode (dado que no es una medida, sino más bien una designación), solo que el método la ignoraba por considerarla que contenía números, por lo cúal se excluyó de la transformación.
    - Según se nos sugiere, debemos trabajar con menos registros por una cuestión de espacio de la herramienta utilizada. Nos indica recortar los registros a 7447.
    - Consideramos a los fines de este entregable continuar con lo sugerido pero sabiendo que en un caso real lo conveniente sería trabajar con menos columnas para que esto no suceda y dejar tanta cantidad de datos afuera

- Numéricas
    1. Rooms: Cantidad de habitaciones
    2. Distance: Distancia al centro de la ciudad
    3. Price: Precio de la propiedad. La variable a predecir
    4. Bathroom: Número de baños de la propiedad
    5. Car: Número de espacios para auto en la propiedad.
    6. Landsize: Tamaño del terreno de la propiedad. En m2

7. BuildingArea: Tamaño de la construcción de la propiedad. En m2
8. YearBuilt: Año en el que fue construida la propiedad
9. Propertycount: Número de propietarios que existe en el barrio. Entendemos que hace referencia a la cantidad de viviendas ocupadas. Cuán poblado sea el barrio también influye en su precio.
10. Price AirBnb: Se agrega la mediana del precio de publicaciones de la plataforma AirBnB en el mismo código postal.
11. Weekly_price: Se agrega la mediana del precio semanal de publicaciones de la plataforma AirBnB en el mismo código postal
12. Monthly_price*:* Se agrega la mediana del precio mensual de publicaciones de la plataforma AirBnB en el mismo código postal.

Transformaciones

- Todas las características numéricas fueron escaladas.
- La columna `Suburb` fue imputada primero con la información de la columna Council Area del mismo dataset de Melbourne, y luego para los valores que aún quedaban nulos, a partir de la información de la columna City del dataset de AirBnB.
- Las columnas `YearBuilt` y 'BuildingArea fueron imputadas utilizando el algoritmo KNeighborsRegressor.

Datos aumentados

Se agregan las 18 primeras columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado. Tomamos m = 18 porque si bien el gráfico nos muestra un primer y gran quiebre del ratio de explicación de la varianza en m = 1, nos parece muy poca cantidad de columnas a considerar, además sólo nos estaríamos quedando con un 15% de la información. En un segundo quiebre y con 18 variables conservamos un 66% de la misma.