Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Exploración y Curación de Datos

Entregable: Parte 2

Ejercicio 5: Documentación

El presente Reporte muestra las operaciones realizadas sobre dos conjuntos de datos con los que hemos trabajado:

1) Datos de una competencia Kaggle sobre estimación de precios de ventas de propiedades en Melbourne, Australia;
2) Información adicional respectiva al entorno de una propiedad a partir del conjunto de datos de AirBnB.

SELECCIÓN DE VARIABLES

A continuación se listan los atributos del Dataset N°1 que NO se consideran relevantes para definir el precio:

- Address: para este estudio no se considera la dirección, se considerara el barrio para el precio.
- SellerG: para este estudio no se considerara al vendedor.
- Date: se considera que la fecha de venta no debería incidir en el precio.
- Bedroom2: para este estudio no se considera, se tiene en cuenta Rooms.
- Lattitude: para este estudio no se considera, se considerara el barrio para el precio.
- Longtitude: para este estudio no se considera, se considerara el barrio para el precio.
- Regionname: para este estudio no se considera, se considerara el barrio para el precio.
- Propertycount: para este estudio no se considera, se considerara el barrio para el precio.
- Postcode: Es importante para combinar los Dataset pero no para predecir el precio de las viviendas.
- CouncilArea: No lo consideramos en nuestro análisis.
- Landsize: No lo consideramos en nuestro análisis por tener muchos ceros.
- Method: No consideramos la variable en nuestro análisis.
- Distance: No se considera la distancia ya que se considera el barrio.

Los atributos que consideramos importantes para la predicción del precio de una propiedad son:

Características categóricas

- Suburb: Region.
- Type: br - bedroom(s); h - house, cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.

Características numéricas

- Rooms: Número de habitaciones.
- Price: Precio en dólares.
- Bathroom: Número de baños.
- Car: Números de cocheras.
- BuildingArea y YearBuilt: Estas dos variables (Superficie construida y año de construcción) son relevantes para predecir el precio de la vivienda y por ende aun teniendo muchos registros "nulos" no nos anticipamos a descartarlas de nuestro análisis, sino por el contrario aplicaremos técnicas de imputación.

Atributos del conjunto de datos elaborado a partir de datos de la plataforma Airbnb, Dataset N°2:

1. 'AirB_ price_mean' (Precio promedio diario de alquiler agrupado por código postal).

2. 'AirB _price_min' (Precio mínimo diario de alquiler agrupado por código postal).

3. 'AirB _price_max' (Precio máximo diario de alquiler agrupado por código postal).

4. 'AirB _price_median' (Precio diario correspondiente a la mediana de alquiler agrupado por código postal).

CRITERIOS DE EXCLUSIÓN

- Eliminación de outliers en la base de datos correspondiente al dataset N°1 (melb_data.csv):

1. Atributo ’YearBuilt’: se eliminaron valores extremos menores a 1900.

2. Atributo ’BuidingArea’: se eliminaron valores mayores a 2000.

3. Atributo ’Price’: se eliminaron valores menores al percentil 5% y mayores al percentil 95% (400.000 – 2.223.800).

4. Atributo ‘Suburb’: Realizamos un conteo de frecuencia de propiedades según su Barrio y consideramos que debe existir al menos 40 propiedades por barrio para que la muestra sea representativa.

- Utilizamos la variable zipcode para unir los conjuntos de datos. Del Dataset N°2 sólo incluimos los zipcode que tengan al menos 30 registros, con el objetivo de que la información agregada sea representativa.

TRANSFORMACIONES

• La variable ‘Bathroom’ fue imputada con valor 1 (uno) para aquellos casos en los que figuraba con 0 (cero).

• Todas las características categóricas fueron codificadas utilizando el método “OneHotEncoder”.

• Todas las características numéricas fueron estandarizadas utilizando el método “RobustScaler” que es más robusto ante la existencia de valores atípicos que StandarScaler.

• Las columnas ’YearBuilt’ y ’BuidingArea’ fueron imputadas aplicando una instancia de IterativeImputer con un estimador KNeighborsRegressor. También aplicamos a modo ilustrativo el estimador: BayesianRidge.

• Las variables 'Car', 'airbnb_price_min', 'airbnb_price_median', 'airbnb_price_mean', 'airbnb_price_max' también fueron imputadas aplicando una instancia de IterativeImputer con un estimador KNeighborsRegressor

DATOS AUMENTADOS

• Se agregan las primeras ocho columnas obtenidas a través del método de PCA. Estas 8 componentes explican el 90 % del total de datos.