Operaciones para obtener el Conjunto de Datos Final

AEyCD - Grupo 32

Reporte

Garay, Carolina del Valle
Ormaechea, Sebastián Gabriel
Ramos, Pablo Nicolás

El presente Reporte muestra las operaciones realizadas a dos dataframes que fueron unidos. Dicho conjunto de datos corresponden a distintos atributos de las propiedades de la ciudad de Melbourne, Australia. Ambos dataset se encuentran disponibles en `https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/melb_data.csv` **y** `https://drive.google.com/file/d/1sUR2Odse85vmQn62yYaEsaUCRJo28Nj2/view?usp=sharing`**. Se seleccionaron los atributos que pueden contribuir a la predicción del precio de la propiedad, detallados en las secciones siguientes.**

1 Criterios de exclusión

- Eliminación de outliers en la base de datos correspondiente al dataset "melb_data.csv":
  1. Atributo 'YearBuilt': se eliminó el valor extremo 1196.
  2. Atributo 'BuidingArea': se eliminaron valores <10 y $\geq$792.
  3. Atributo 'Car': se consideraron valores de 0 a 6, descartando missing values.
  4. Atributo 'Rooms': se eliminó el valor extremo 10.
  5. Atributo 'Price': se eliminaron valores < 200000 y > 3000000.
- Seleccionamos del conjunto de datos elaborado a partir de datos de la plataforma Airbnb los zipcodes que tenían al menos 3 registros.

2 Características seleccionadas

Las siguientes variables numéricas y no numéricas están referidas a la propiedad.

2.1 Características categóricas

- Atributo 'Regionname' (Nombre de la región en la que se ubica la propiedad): 8 valores posibles
- Atributo 'Type' (tipo de propiedad): 3 valores posibles

Todas las características categóricas fueron codificadas mediante la creación de instancias del objeto OHE de codificadores de categoría. Se utlizó el parámetro "$use\_cat\_names = True$", el cual permite que los valores de categoría se incluyan en los nombres de columna frecuente.

2.2 Características numéricas

- Atributos del dataset "melb_data.csv"
  1. 'YearBuilt' (Año de construcción)
  2. 'BuildingArea' (Área ocupada)
  3. 'Car' (Cantidad de cocheras por propiedad)
  4. 'Price' (Precio de la propiedad)

5. 'Rooms' (Cantidad de habitaciones)

- Atributos del conjunto de datos elaborado a partir de datos de la plataforma Airbnb
  1. 'AirB_daily_price_mean' (Precio promedio diario de alquiler)
  2. 'AirB_weekly_price_mean' (Precio promedio semanal de alquiler)
  3. 'AirB_monthly_price_mean (Precio promedio mensual de alquiler)

3 Transformaciones

- Todas las características categóricas codificadas y las numéricas fueron estandarizadas
- Las columnas 'YearBuilt' y 'BuidingArea' fueron imputadas aplicando una instancia de IterativeImputer con un estimador KNeighborsRegressor
- Las columnas 'AirB_daily_price_mean', 'AirB_weekly- _price_mean' y 'AirB_monthly_price_mean' fueron imputadas aplicando una instancia de IterativeImputer con un estimador KNeighborsRegressor

4 Datos aumentados

- Se agregan las 8 primeras columnas obtenidas a través del método de PCA. Estas 8 componentes resultan suficientes para explicar el 95 % del total de datos.