Práctico de Exploración y Curación de Datos – Grupo 5

Análisis de los precios de las casas en Melbourne

El dataset inicial de este trabajo proviene de un conjunto de datos de Kaggle[1] sobre estimación de precios de ventas de propiedades en Melbourne, Australia. En particular, se utilizó un conjunto de datos reducido producido por DanB[2] y facilitado en un servidor de la Universidad Nacional de Córdoba[3].

El set de datos contiene 21 variables y 13.580 observaciones. A continuación, se describen algunas variables, tal y como están definidas en el sitio de Kaggle.

Suburb	
Address	
Rooms	Number of rooms
Type	br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
Price	Price in dollars
Method	S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available
SellerG	Real Estate Agent
Date	Date sold
Distance	Distance from CBD
Postcode	
Bedroom2	Scraped # of Bedrooms (from different source)
Bathroom	Number of Bathrooms
Car	Number of carspots
Landsize	Land Size
BuildingArea	Building Size
YearBuilt	
CouncilArea	Governing council for the area
Lattitude	
Longitude	
Regionname	General Region (West, North West, North, North east …etc)
Propertycount	Number of properties that exist in the suburb.

[1] https://www.kaggle.com/dansbecker/melbourne-housing-snapshot

[2] https://www.kaggle.com/dansbecker

[3] https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/melb_data.csv

Transformaciones realizadas en el entregable 1

1- Limpieza de datos

En primer lugar, se trabajó con los outliers. Para ello se tomaron las estadísticas descriptivas estándar y se evaluó qué variables contenían una alta dispersión. Después se definieron límites superiores para las distintas variables, y se eliminaron las observaciones que estuvieran fuera de estos límites. A continuación, se presentan los límites definidos y sus justificaciones:

Variable	Min	Med	Máx	Límite	Justificación
Rooms	1	3	10	8	Se limitará a 8 rooms, dado que es el valor más alto que se presenta razonable. Considerando que son "espacios" de la propiedad y no "habitaciones", se entiende que no es un outlier. También se controla que el número de habitaciones sea mayor a 0.
Bedroom2	0	3	20	Se elimina	La variable "Bedroom2" debería limitarse a 8 como máximo, para que no supere el valor de "Rooms". Sin embargo, las variables "Rooms" y "Bedroom2" se encuentran altamente correlacionadas, por lo que esta última será descartada
Bathroom	0	1	8	5	Considerando que la variable "Rooms" fue limitada a 8, la cantidad de baños debería ser menor a este valor. Por esto, la variable fue limitada a 5 o menos, ya que no parece razonable que haya una proporción tan alta de baños en una casa. También se imputa como 1 cuando esta variable es 0.
Car	0	2	10	6	La variable "Car" es limitada a 6. La frecuencia de esta variable para una cantidad mayor a 6 es demasiado baja en comparación con la cantidad total de datos, por lo éstos serán tratados como outliers.
BuildingArea	0	126	44515	500	Por la distribución de la variable, se toma 500 como límite, por contener más del 99% de las observaciones.
Landsize	0	440	433014	1500	Ya que más del 99% de las observaciones cuentan con un Landsize menor a 1500, se tomará este como límite
YearBuilt	1196	1970	2018	<1196	Se dejará afuera el outlier "1196" de la variable "YearBuilt", entendiendo que se trata de un error por tener cerca de 825 años de antigüedad.

Luego de trabajar con los outliers, se definió con que columnas trabajar. El objetivo era dejar solamente aquellas columnas que aportaran información importante para la predicción del precio, evitando dejar variables con información redundante.

De las variables geográficas, "Address", "Distance", "Regionname", "Longitude", "Lattitude", “CouncilArea” y ”Suburb” se dejarán solo “CouncilArea” y “Suburb”, ya que todas brindan información similar, pero “CouncilArea” lo hace de una manera ni muy agregada como “Regionname” ni muy desagregada como “Suburb”, y “Suburb” se mantiene porque es necesaria para imputar los datos faltantes en “CouncilArea”.

Adicionalmente, se elimina la variable “Bedroom2” por estar fuertemente relacionada con la variable “Rooms”.

2- Información de AirBnb

Buscando enriquecer el conjunto de datos se tomó información relativa al entorno de una propiedad originaria del sitio AirBnb[4]. Estos nuevos datos se unieron al dataset de Melbourne a través del Código Postal. La única variable que se agregó de este dataset es el promedio de la variable price, de tal manera que hubiera una relación biunívoca entre ambos conjuntos de datos. El motivo por el que solo se tomó esta variable es porque las demás variables del nuevo dataset respecto al entorno de las propiedades eran mayormente geográficas, proveyendo datos redundantes a los ya considerados en nuestro dataset.

3- Imputación

En un primer momento se imputa la variable “CouncilArea” según los datos de “Suburb” siguiendo la regla de mayor frecuencia. Adicionalmente se imputan los pocos datos faltantes del precio según AirBnb, de acuerdo con el Postcode.

Transformaciones realizadas en el entregable 2

4- Algunas transformaciones adicionales

Como se mencionó anteriormente, Suburb no se va a trabajar ya que cuenta con más de 300 valores únicos, lo que complica su manejo al hacer encoding, es por ello que se elimina del dataset.

De CouncilArea, que, si se mantiene, se unifican todas las observaciones que pertenezcan a CouncilAreas con menos de 50 observaciones, en un CouncilArea “Other”.

En cuanto a los vendedores, a quienes habían vendido menos de 40 propiedades se los agrupó en 3 conjuntos: todos aquellos que solo hubieran vendido hasta 10 unidades se los clasificó como “Casual”, entre 11 y 20 inclusive, como “Medium” y entre 21 y 40 inclusive como “Upper Medium”. De esta manera estas categorías quedaron con 506, 357 y 534 observaciones, respectivamente.

5- Encoding

Para poder trabajar con las variables categóricas se procedió a realizar el encoding de las mismas, previa eliminación de la columna de fecha, mediante la funcion OneHotEncoder.

6- Imputación

La imputación por KNN de los valores faltantes en las columnas “YearBuilt” y “BuildingArea” se realizó dos veces, en la primera, solo se tuvieron en cuenta los valores de “YearBuilt” y de “BuldingArea”, y en la segunda se usaron todas las variables disponibles en el dataset.

7- PCA

Luego de haber realizado la imputación, se procedió a estandarizar las variables (excluyendo la variable precio, por ser variable objetivo) y se realizó el análisis de componentes principales. Del elbowplot se definió que se trabajaría con las 8 primeras componentes, y se anexaron las mismas

[4] https://www.kaggle.com/tylerx/melbourne-airbnb-open-data?select=cleansed_listings_dec18.csv

al dataset original sin estandarizar, pero con los datos faltantes ya imputados según la información proveniente de todo el dataset.

8- Composición del resultado y exportación del dataset

Como al aplicar OneHotEncoding ya se había aplicado una función que permitía mantener los nombres de las columnas, no fue necesario repetir este paso y directamente se pasó a la exportación del dataset.