Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Exploración y Curación de Datos

Edición 2021

Trabajo práctico entregable

Documentación

Descripción de columnas

- **Suburb**: Suburb of property inside Council Area
- **Address**: Address of property
- **Rooms**: Number of rooms in house including bedrooms
- **Price**: Price in US 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    - W - withdrawn prior to auction;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.
- **Type**:
    - br - bedroom(s);
    - h - house,cottage, villa, semi, terrace;
    - u - unit, duplex;
    - t - townhouse;
    - dev site - development site;
    - res - other residential.
- **SellerG**: Real State Agent (Agente inmobiliario)
- **Date**: Date sold
- **Distance**: Distance from CBD in Kilometres (Central Business District)
- **Post Code**: Código postal.

- **Regionname**: General Region (West, North West, North, North east …etc)
- **Propertycount**: Number of properties that exist in the suburb.
- **Bedroom2** : Scraped # of Bedrooms (from different source)
- **Bathroom**: Number of Bathrooms
- **Car**: Number of carspots (Número de plazas de coche)
- **Landsize**: Land size in Metres (Tamaño del terreno)
- **BuildingArea**: Building Size in Metres (Tamaño área construida)
- **YearBuilt**: Self explanitory
- **CouncilArea**: Governing council for the area (Consejo de gobierno de la zona)
- **Lattitude**: Self explanitory
- **Longtitude**: Self explanitory

Criterios de exclusión de columnas

Para incluir columnas se tuvo en cuenta el objetivo: 'Predicción del valor de la propiedad'. Por ello se descartaron las siguientes columnas:

Numericas:

- Address
- Lattitude
- Longtitude
- Bedroom2
- Bathroom
- BuildingArea*
- YearBuilt*

Categoricas:

- CouncilArea*
- Method
- SellerG
- Date
- Postcode

* Se mantienen para resolver ejercicios de imputación pero según los análisis realizados se descartarían para predecir el precio de las propiedades.

Valores atípicos y eliminación de outliers para variables numéricas

Rooms

- Mantenemos los casos donde hay 5 o menos habitaciones

```
melb_df = melb_df[(melb_df['Rooms'] <= 5]
```

Car

- Incluímos sólo los casos donde hay 6 plazas o menos.

```
melb_df = melb_df[(melb_df["Car"] <= 6) | (melb_df["Car"].isna())]
```

Landsize

- Los valores de la variable se convierten de string a float.
- Convertimos los valores en 0 a nulos.
- Mantenemos valores menores a 433013

```
melb_df = melb_df[melb_df['Landsize'] < 433013]
```

```
melb_df['Landsize'] = melb_df['Landsize'].replace(0, numpy.nan)
```

Price

- Mantenemos valores iguales o menores al valor extremo 7874734.

```
price_melb_df = melb_df[melb_df["Price"] <= 7874734]
```

Distance

- Conservamos la variable original.

YearBuilt

- Eliminamos un outlier extremo inferior: 1196.

```
melb_df = melb_df[melb_df['YearBuilt'] > 1196]
```

BuildingArea

- Decidimos eliminar el valor atípico extremo igual a 44515.
- Reemplazamos los valores en 0 por nulos.

```
melb_df = melb_df[melb_df['BuildingArea'] < 44515]
```

```
melb_df['BuildingArea'] = melb_df['BuildingArea'].replace(0, numpy.nan)
```

Tratamiento de variables categóricas:

Type

- Conservamos la variable original.

Suburb

- Reagrupamos los 314 suburbios en 10 grupos según la mediana del precio.

Regionname

- Reagrupamos las categorías 'Eastern Victoria', 'Northern Victoria' y 'Western Victoria' en 'Victoria'.
- Reagrupamos las categorías 'Eastern Metropolitan' y 'South-Eastern Metropolitan' en 'Eastern Metropolitan'.

CouncilArea

Imputamos los valores faltantes basándonos en la variable `suburb` y complementando con un dataset que pertenece al gobierno de Melbourne: https://raw.githubusercontent.com/Natali-PP/diplodatos2021/main/EyCD/victoria-councils-transformed.txt

Agregaciones

Se realiza un merge con el dataframe de AirBnB: 'https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/cleansed_listings_dec18.csv

Para mergear usamos la columna 'zipcode' del dataframe original con 'Postcode' del dataframe de Airbnb.

Del dataframe de AirBnB, calculamos la media por zipcode de las siguientes variables que decidimos agregar a nuestro set de datos:

- airbnb_price_mean
- airbnb_weekly_price_mean
- airbnb_monthly_price_mean

Transformaciones

Imputaciones con IterativeImputer

- Luego de la agregación de datos de AirBnB, imputamos los nulos de las columnas `Car`, `Landsize`, `airbnb_weekly_price_mean` y `airbnb_monthly_price_mean`, usando IterativeImputer con el estimador por defecto (BayesianRidge).

Encoding

- Se realiza en Encoding de todas las variables categóricas.
- Para ellos se utiliza `DictVectorizer` que combina todas las columnas categóricas del encoding con los datos numéricos existentes.

Para poder realizar el encoding se hicieron las siguientes modificaciones previas:

- Se renombran las columnas de Airbnb:
  - airbnb_price_mean --> PriceMeanAirbnb
  - airbnb_weekly_price_mean --> PriceWeeklyMeanAirbnb
  - airbnb_monthly_price_mean --> PriceMonthlyMeanAirbnb
- Se cambia el orden de las columnas para que estén las numéricas primero, y despues las categóricas.
- Se eliminan (por el momento) las columnas `YearBuilt` y `BuildingArea`.

Imputaciones con KNN

- Se imputan los valores faltantes de las variables `YearBuilt` y `BuildingArea` usando `IterativeImputer` con estimador `KNeighborsRegressor` y tomando como parámetro todas las variables numéricas del dataset.

Para poder realizar las imputaciones se hicieron las siguientes modificaciones previas:

- Se transforma la matriz de esparsa a densa (Tamaño= 1,34 MB).
- Se agregan las columnas `YearBuilt`, `BuildingArea` usando `numpy.hstack`.
- Se separan las columnas numéricas de las categóricas.
- Se escalan entre -1 y 1 los datos numericos usando `StandardScaler`.

Reducción de dimensionalidad

- Se aplica `sklearn.decomposition.PCA` para calcular n=20 componentes principales.
- Se eligen los 4 primeros componentes del resultado por concentrar más del 60% de la variabilidad total explicada.

Datos aumentados

- Se agregan al dataset las 4 primeras componentes obtenidas a través del método de PCA.
- El dataset final consta de las siguientes columnas:
  - 'YearBuilt'

- 'BuildingArea'
- 'Car'
- 'Distance'
- 'Landsize'
- 'Price'
- 'Rooms'
- 'PriceMonthlyMeanAirbnb'
- 'PriceMeanAirbnb'
- 'PriceWeeklyMeanAirbnb'
- 'Suburb_cat=Suburb_group1'
- 'Suburb_cat=Suburb_group10'
- 'Suburb_cat=Suburb_group2'
- 'Suburb_cat=Suburb_group3'
- 'Suburb_cat=Suburb_group4'
- 'Suburb_cat=Suburb_group5'
- 'Suburb_cat=Suburb_group6'
- 'Suburb_cat=Suburb_group7'
- 'Suburb_cat=Suburb_group8'
- 'Suburb_cat=Suburb_group9'
- 'Type=h'
- 'Type=t'
- 'Type=u'
- 'reg_cat=Eastern Metropolitan'
- 'reg_cat=Northern Metropolitan'
- 'reg_cat=Southern Metropolitan'
- 'reg_cat=Victoria'
- 'reg_cat=Western Metropolitan'
- 'pca1'
- 'pca2'
- 'pca3'
- 'pca4'

FIN