Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Exploración y Curación de Datos

Edición 2022

Entregable 2 - Parte 2 - Ejercicio 5

Grupo 27:
Guillermo Alonso
María Eugenia Bernaschini
Juan Cruz Bordón
Javier Carabajal

Documentación

Objetivo general

Procesar la información necesaria para estimar los precios de ventas de propiedades en Melbourne, Australia de la base de datos de la competencia Kaggle (reducido por DanB) y de la base de datos de AirBnB.

Objetivos específicos

- Selección de variables relevantes.
- Eliminación de valores extremos.
- Combinación de bases de datos.
- Codificación de variables categóricas.
- Imputación de valores perdidos.
- Cálculo y selección de componentes principales.
- Armado del dataset procesado.

Criterios de exclusión

- Para el problema de predicción de valores de las propiedades excluimos del análisis las siguientes variables de la base de ventas de propiedades en Melbourne, Australia:

  Address: si bien tiene que ver con la ubicación del inmueble (dirección), es una variable difícil de trabajar por su gran variabilidad.
  Method: tiene que ver con la forma en la que fue vendida la propiedad.
  SellerG: agencia que llevó a cabo la venta.
  CouncilArea: área de gobierno local. Esta variable no fue seleccionada ya que está relacionada con la variable Suburb.

- Los precios de la base de Melbourne superiores a 4003420 dólares (cuantil 0,996) fueron considerados como valores atípicos y eliminados de la base.

- En cuanto a la variable Suburb de la base de Melbourne, se tomaron aquellos suburbios con frecuencia mayor o igual a 10, ya que por debajo de esa frecuencia se consideraron poco representados.

- En la variable Price de la base AirBnB se eliminaron valores extremos (superiores a 1501 dólares, que representa el cuantil 0,998).

- Para la variable Zipcode de la base AirBnB se tomaron aquellos valores con frecuencias mayores o iguales a 9, ya que a partir de esa frecuencia creemos que son relevantes.

Características seleccionadas

Características categóricas

- *Date:* fecha de venta. 58 valores posibles.
- *Type:* tipo de propiedad. 3 valores posibles.
- *Regionname:* nombre de la región general (Norte, Sur, etc.). 7 valores posibles.
- *Suburb:* zona residencial en donde se encuentra ubicada la propiedad. 212 valores posibles.

Todas las características categóricas fueron codificadas con un método **OneHotEncoding.** Para la variable Suburb se utilizaron los 50 valores más frecuentes.

Características numéricas

- *Rooms:* cantidad de ambientes que tiene la casa.
- *Price:* precio al que fue vendida la propiedad.
- *Distance:* distancia al distrito financiero.
- *Postcode:* código postal que usaremos para combinar las bases y tiene que ver con la ubicación de la propiedad.
- *Bedroom2:* cantidad de habitaciones.
- *Bathroom:* cantidad de baños.
- *Car:* cantidad de cocheras.
- *Landsize:* tamaño del terreno.
- *BuildingArea:* tamaño del terreno edificado.
- *YearBuilt:* año de construcción de la propiedad.
- *Lattitude:* latitud, tiene que ver con la ubicación de la propiedad.
- *Longtitude:* longitud, tiene que ver con la ubicación de la propiedad.
- *Propertycount:* cantidad de propiedades que hay en el suburb.
- *airbnb_price_median:* precio mediano diario de publicaciones de la plataforma AirBnB por código postal.
- *airbnb_weekly_price_median:* precio mediano semanal de publicaciones de la plataforma AirBnB por código postal.
- *airbnb_monthly_price_median:* precio mediano mensual de publicaciones de la plataforma AirBnB por código postal.
- *airbnb_record_count:* conteo de datos por cada zipcode de la base AirBnB.

Transformaciones

- Para la base AirBnB se construyeron cuatro nuevas variables: *airbnb_price_median* (precio mediano diario de publicaciones de la plataforma AirBnB por código postal), *airbnb_weekly_price_median* (precio mediano semanal de publicaciones de la plataforma AirBnB por código postal), *airbnb_monthly_price_median* (precio mediano mensual de publicaciones de la plataforma AirBnB por código postal) y *airbnb_record_count* (conteo de datos por cada zipcode de la base AirBnB). Se eligió la mediana ya que las distribuciones de precios son asimétricas a derecha, por lo tanto la mediana es una mejor medida de posición central en comparación a la media. Luego se combinaron las bases de Melbourne con la base AirBnB por código postal (ver base en repositorio).

- Las columnas YearBuilt, BuildingArea, car, airbnb_price_median, airbnb_weekly_price_median, airbnb_monthly_price_median y airbnb_record_count fueron imputadas utilizando el método **IterativeImputer** con un estimador **KNeighborsRegressor**. Para implementar las imputaciones escalamos al rango [0,1] y probamos incluir, por un lado, todas las columnas, por otro lado, sólo aquellas columnas con valores pérdidos y, por último, sólo las columnas YearBuilt y BuildingArea. Luego de realizar un análisis de distribución comparativo decidimos quedarnos con la imputación que utiliza todas las columnas, ya que logró mejores resultados en cuanto al parecido de la distribución de cada variable imputada con la distribución original.

Datos aumentados

Para el cálculo de las n (n=20) componentes principales se escalaron los datos al rango [0,1], luego se aplicó el algoritmo PCA y se eligieron las 5 componentes que mejor explican la variabilidad. Finalmente, dichas componentes se agregaron a la base de datos procesada (ver base en repositorio).