Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Exploración y Curación de Datos

Edición 2022

Grupo 2. Integrantes:

Diego A. Gómez

Natalia A. Kunzmann

Natalia C. Graselli

Patricia V. Gonzalez

M. Virginia Romero Messein

El objetivo de este documento es explicitar las decisiones tomadas en el proceso de exploración y curación de datos.

Se trabajó con dos conjuntos de datos principales, el primero es producido por DanB[1] y corresponde a una competencia Kaggle[2] acerca de la estimación de precios de ventas de propiedades en Melbourne, Australia. El segundo conjunto proviene de Airbnb[3].

La pregunta que nos hicimos es la siguiente: ¿qué variables pueden ayudar a predecir el precio de una propiedad? Para ello fue necesario hacer una serie de transformaciones a los datos con la finalidad de crear un nuevo dataset que sea apto para entrenar un modelo.

Características seleccionadas interesantes para la predicción de precio:

Características categóricas:

Suburb: Barrio.

Type: Tipo. Hay tres valores posibles: h - casa, cabaña, villa, semi, terraza; u - unidad, dúplex; t - casa adosada.

Características numéricas:

Rooms: Habitaciones: Número de habitaciones.

Bathroom: Baño. Número de Baños.

Bedroom 2: Dormitorio 2: número de dormitorios.

Price: Precio en dolares.

---

[1] DanB | Grandmaster | Kaggle

[2] Melbourne Housing Snapshot | Kaggle

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

BuildingArea: Tamaño del edificio.

LandSize: Tamaño del terreno.

YearBuilt: Año de construcción.

Postcode: Código postal.

Criterios de selección de variables:

- *Suburbio*: por considerar que no es lo mismo comprar una propiedad en un barrio que en otro.
- *Rooms, Bathroom y Bedroom 2*: por considerar que el tamaño de una propiedad, incide en el número habitaciones y baños tenga, y por lo tanto, mayor será su precio.
- *Type:* por entender que los distintos tipos de propiedades pueden llegar a influir en el precio.
- *Precio*: porque es la columna principal a tener en cuenta para la predicción de un precio.
- *Land Size y Building Area*: porque son útiles para explorar cómo incide el tamaño del terreno y la superficie edificada en el precio.
- *Year Built*: porque se presupone una relación entre la antigüedad de la propiedad y su precio.
- *Postcode*: porque será útil para poder hacer el merge con el df de Airbnb a través de este valor.

Filtrado de valores extremos

Para filtrar datos extremos, se tomó como límite el percentil 99.9 de cada variable numérica.

Se descubrió que filtrando de esta manera quedarían aprox el 50% de los valores (inicial 13580, luego de filtro 6795) únicamente. Luego de analizar los resultados, se observa que esto se debe a que gran parte de los valores de *Building Area* y *Year Built* están vacíos por lo que se decide proceder a no tocar la columna Building Area por el momento.

Método 2:

-Se eliminó el valor extremo de *Landsize* de 433.014.

-Se eliminó el valor extremo de *BuildingArea* de 44.515.

-Se elimina el valor extremo de *Rooms* de 10.

-Se eliminaron los valores altos de *Bathroom* de 7 y 8.
-Se eliminó el valor extremo de *Bedroom* de 20 y valores altos de 9 y 10.
-Se eliminó el valor extremo de *Price* de 9.000.000.
-Se eliminó el valor extremo de *YearBuilt* de 1196.

Transformaciones:

Primero transformamos las variables categóricas en numéricas. Comenzamos quitando las columnas que no nos interesaban en este apartado, éstas son: *YearBuilt* y *Building Area*.

"One-Hot Encoding" es un método de sklearn mediante el cual convertimos datos categóricos en numéricos, para después poder aplicarlos a un modelo de ML. De esta manera las variables categóricas se convierten en variables binarias.

El resultado es una matriz esparsa, a la cual le agregamos nuevamente las variables numéricas que habíamos descartado para el One Hot Encoding.

Por otro lado, para lograr la estandarización, utilizamos el método "StandasScaler" de sklearn (podríamos también haber usado otros, como MinMaxScaler).

Por último, todos los valores faltantes se imputaron utilizando el método "IterativeImputer"

Datos aumentados

Se agregan las primeras 4 columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado *pca1_scaled, pca2_scaled, pca3_scaled, pca4_scaled.*

Repositorio:

Parte1:
https://colab.research.google.com/drive/1jcsUDUEJVAp33096b7qGOIZvKElsd-Mo#scrollTo=WekVJ9bwCQsO

Parte2:
https://colab.research.google.com/drive/1DhohjPTeNrdYUc3KBXfB07oLXhjv29xN#scrollTo=Iikrm1cVFMMB