Documentación

Criterios de exclusión de ejemplos

Excluimos las siguientes variables ya que son para identificar cada fila de datos: id, listing_url, scrape_id, name, host_id, host_url, host_name, picture_url, host_location,host_thumbnail_url, host_picture_url,host_has_profile_pic

Las siguientes columnas fueron excluidas porque no aportan información útil para la predecir el precio de venta: summary, space, description, neighborhood_overview, notes, transit, access, interaction, house_rules, host_about.

Las siguientes columnas también fueron excluidas porque solo podrían impactar indirectamente al precio de alquiler, (ya que se refieren al anfitrión) confianza: host_since, host_response_time,host_response_rate, host_is_superhost, host_neighborhood, host_verifications, host_identity_verified,calculated_host_listings_count.

Excluimos country_code, country porque no aportan información relevante
Descartamos weekly_price y monthly_price ya que ya verificamos que tenían una cantidad elevada de valores nulos

Las siguientes columnas fueron excluidas porque son más bien informativas y no ayudan a predecir el precio de las propiedades: is_location_exact, calendar_last_scraped, first_review, last_review, has_availability, availability_30, availability_60, availability_90, availability_365, calendar_updated, requires_license,license,instant_bookable, cancellation_policy, require_guest_profile_picture require_guest_phone_verification.

Las siguientes columnas posiblemente podrían ser usadas para predecir el precio de alquiler, pero no las consideramos para predecir el precio de venta de propiedades en Melbourne: last_scraped, Street (difícil de codificar y menos informativa que el barrio), review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin,review_scores_communication, review_scores_location, review_scores_value,number_of_reviews,reviews_per_month,property_type , room_type, accommodates, bathrooms, bedrooms, beds, bed_type, amenities, security_deposit, cleaning_fee, guests_included, extra_people, minimum_nights, maximum_nights.

Eliminando los valores del precio de alquiler superiores a $500 para trabajar con valores medidas de tendencia central más representativas. También eliminamos los iguales a 0 y 1

Del dataframe orginal estamos dejando fuera 211 filas de datos, que representan menos del 1% de los datos.

Quitamos los 62 registros que no tenían información de la cochera para poder trabajar con las metodologías de Encoding.

Interpretación de las columnas presentes

Exploramos las siguientes variables como alternativas a ser usadas para cruzar los datos de las dos bases:
zipcode: con PostalCode como se hizo en clase.
latitude y longitude: pero la descartamos por no coincidir exactamente en las dos tablas.
suburb: con Suburb
neighborhood con Suburb?
city / state deberíamos chequear que haya alguna similar en Melbourne, o transformarla con información adicional: ejemplo la ciudad y el estado en la base de Melbourne podrían deducirse del barrio. En este sentido, podrían servir para cruzar pero necesitamos trabajar un poco más los datos.
smart_location: no parece haber ninguna equivalente en la tabla de Melbourne.

Decidimos que las columnas que pueden considerarse relevantes para predecir el precio de venta de la tabla de airbnb son: neighborhood,city, suburb, state, zipcode, latitude, longitude, price.

Datos aumentados

Utilizamos dos bases de datos, una con datos de Melburne y la otra con datos de Airbnb, y creamos un nuevo set de datos para mejorar el modelo.
La unión se realiza a partir de las variables Postcode (de la B.D. Melburne) y Zipcode (de la B.D. Airbnb)
Los datos fueron ingestados utilizando el método: .to_sql('', con=engine, if_exists="replace")

Se construyo una tabla con la variable price transformada

Añadimos las 2 primeras columnas que poseen mayor varianza y podrían explicar mejor los datos.les categóricas y numéricas.

Para la reducción de dimensionalidad estandarizamos las variables

Se construyo una tabla con la variable price transformada

Añadimos las 2 primeras columnas que poseen mayor varianza y podrían explicar mejor los datos.

Características seleccionadas

Al combinar los datasets de ambas tablas, tomamos como columnas relevantes de la tabla airbnb las siguientes:
zipcode
Promedio de price por zipcode
Promedio de weekly_price por zipcode
Promedio de monthly_price por zipcode

Decidimos que las columnas que pueden considerarse relevantes de la tabla de airbnb para predecir el precio de venta son: neighborhood, city, suburb, state, zipcode, latitude, longitude, price.

Se observa que la columna Suburb puede utilizarse para cruzar las bases sin problemas. El porcentaje de registros de la base de Melbourne que tienen un Suburb tambiéen existente en la base de Airbnb es de 99.09% (algo menor que el porcentaje de zipcodes que era de 99.85%).

Una observación es que los códigos postales únicos son menos que los Suburb únicos, lo cual implica que va a haber menos variabilidad (y por lo tanto menos información) en los precios promedio de alquileres que ingestemos.

Elegimos concentrarnos en la variable price, que es la más completa. (porque price, weekly_price, monthly_price, tienen gran cantidad de valores nulos)

Decidimos quedarnos con aquellos zipcodes que tengan al menos 8 registros.

Características categóricas

En el dataset airbnb_df tiene las variables ciudad (city) y otra barrio (neighborhood), y por otra parte en el dataset de melb_df las variables CouncilArea y Suburb.

Se observa que las variables tienen distintos nombres, pero al analizar los datos se observan que categorías son equivalentes. Por lo tanto podemos unir City con CouncilArea, y neighborhood con Suburb

Transformaciones:

El type de las variables Postcode (de la B.D. Melburne) y Zipcode (de la B.D. Airbnb), son diferentes, por lo tanto se cambió el de una de ellas para hacerlas coincidir.

Se seleccionaron las filas y columnas relevantes al problema de predicción de precios de una propiedad e imputaron los valores faltantes de las columnas Suburb y las columnas obtenidas a partir del conjunto de datos airbnb.

Profundizamos el análisis de las columnas viendo nulos y ceros, utilizamos una función status con el método: status( )

Concatenamos los resultados del OneHot Encoding para las variables categóricas y numéricas.

Para la reducción de dimensionalidad estandarizamos las variables