Segunda entrega del proyecto

POR:

Alejandro Herrera Rivera

Juan José Isaza

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2023

Introducción

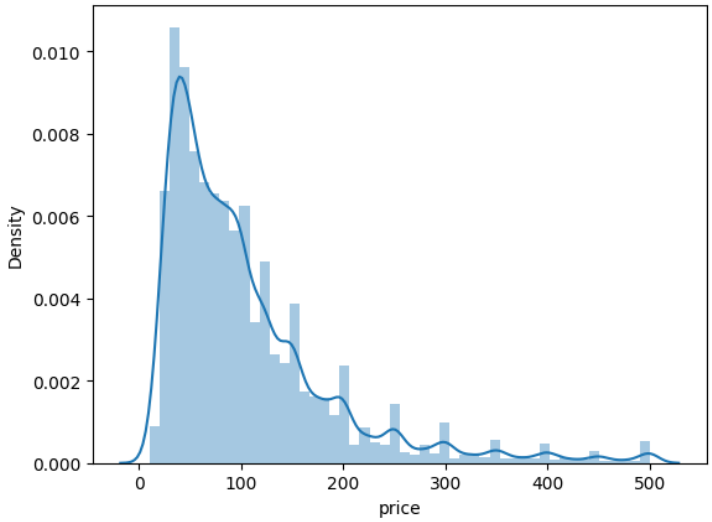
En este trabajo se explorará la implementación de modelos de machine learning para predecir el precio de los Airbnb de la ciudad de Londres, teniendo en cuenta las características de las propiedades así como las referencias ya dadas por los usuarios. De esta manera dar un precio justo que se acomode adecuadamente al valor por el que se debería pagar para la adquisición de este servicio.

Análisis exploratorio

El dataset cuenta con 55284 filas y 42 columnas. Al ver la tabla inicial se pueden ver que algunas variables no se tendrán en cuenta porque son descripciones de las residencias, entre estas encontramos **name, summary, space, description** entre otras. Siendo un total de 11 variables que se eliminarán para facilitar el manejo del dataset.

Variable objetivo:

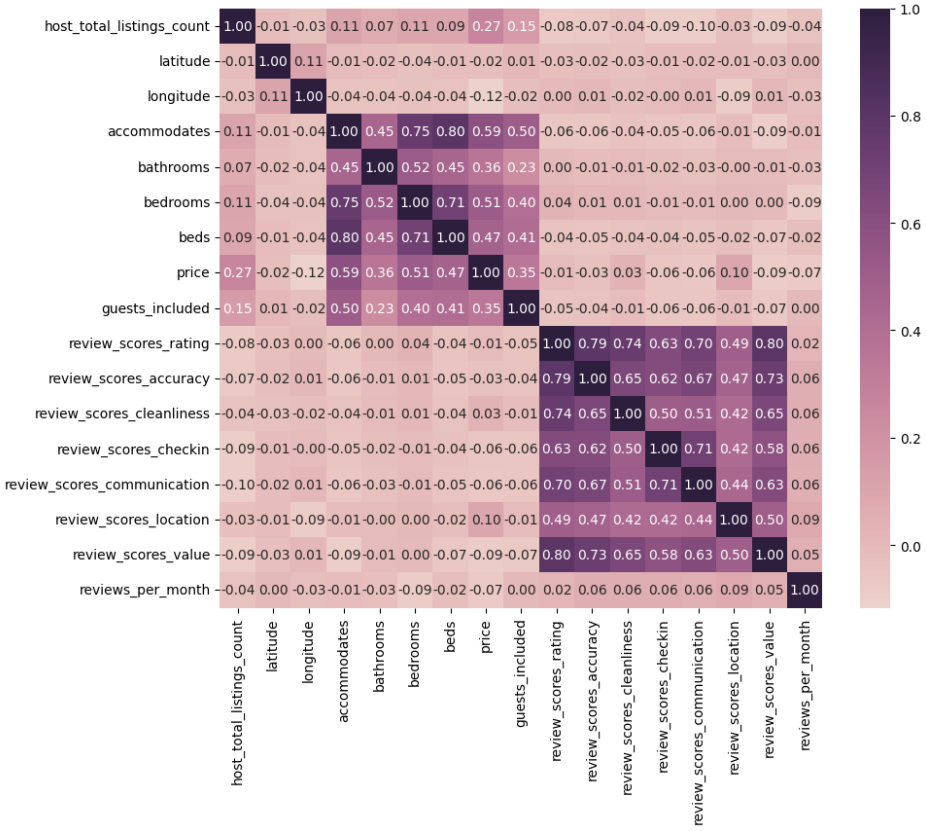
Se realizó un histograma para observar el comportamiento de la variable objetivo, observando que la mayor cantidad de precios se encuentran entre los 50 a los 100 dólares:



Correlaciones:

Se realiza una matriz de correlación entre las variables numéricas, para conocer si

pueden existir problemas de multicolinealidad al haber 2 o más variables independientes altamente relacionadas:



Las variables de reviews de los clientes presentan bastante correlación entre ellas, como por ejemplo **"review scores rating"**, la cual presenta los valores de correlación más altos con las demás variables de reviews, siendo en casi todos los casos mayor a 0.6, al igual que la variable de capacidad "accommodates" se correlaciona mucho con el número de camas **"beds"**, el número de habitaciones "bedrooms" y el precio. Siendo asi **'review_scores_rating'**, **'accommodates'**, **'bedrooms'**, **'review_scores_accuracy'** las cuatro variables que se eliminarán para crear una tabla nueva limpia de esta multicolinealidad.

En cuanto a las variables categóricas se decide aplicar la prueba de chi-cuadrado, seleccionando los resultados mayores a 0.05 los cuales nos indican una alta correlación:

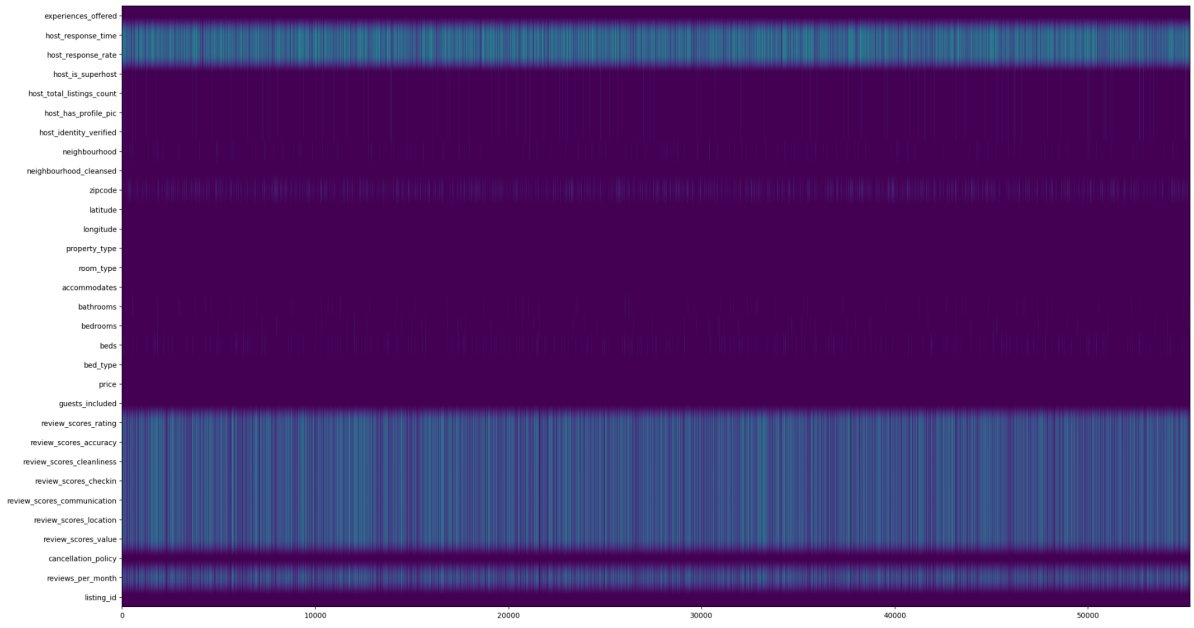
```
('experiences_offered', 'host_has_profile_pic', 0.7262523248759465)
('host_response_time', 'host_has_profile_pic', 0.06519513735833132)
('host_has_profile_pic', 'experiences_offered', 0.7262523248759465)
('host_has_profile_pic', 'host_response_time', 0.06519513735833132)
('host_has_profile_pic', 'neighbourhood', 0.5249147652569358)
('host_has_profile_pic', 'property_type', 0.28057414590348084)
('host_has_profile_pic', 'bed_type', 0.8739667231613887)
('neighbourhood', 'host_has_profile_pic', 0.524914765256935)
('neighbourhood', 'bed_type', 0.06702526850700927)
('property_type', 'host_has_profile_pic', 0.28057414590348084)
('bed_type', 'host_has_profile_pic', 0.8739667231613887)
('bed_type', 'neighbourhood', 0.06702526850700927)
('bed_type', 'cancellation_policy', 0.8928835121337191)
('cancellation_policy', 'bed_type', 0.892883512133719)
```

Las variables **"host has profile_pic"** y **"bed_type"** se correlacionan mucho con otras variables, por lo que se decide eliminarlas para la tabla limpia.

Datos faltantes:

Se observó que **"host_response_rate"** y **"host_response_time"** son las variables que presentan la mayor cantidad de datos faltantes, con un porcentaje de 32.2% cada una y un total de 17802 datos nulos.

Las variables de **'reviews'** son las siguientes con mayor cantidad de nulos, cada una con un porcentaje cercano al 24% de datos faltantes.



Tratamiento de datos

Se eliminan las variables **'listing_id'** y **'zipcode'**, debido a su gran cantidad de

datos únicos, ya que al hacer uso de variables dummies para la generación de modelos estas variables generan más de 80.000 columnas dificultando el uso del dataset.

Nulos:

Ya que no hay variables con una cantidad de datos nulos tan significativa que supere el 50% de sus datos, se decide no eliminar ninguna variable. En este caso se optará por llenar esos datos con la mediana en el caso de las variables numéricas para que los datos no se vean afectados por los datos atípicos que pueden influir la media. Las variables categóricas se llenarán con la moda.

Modelos

Se implementó inicialmente un modelo de Regresión lineal por su fácil implementación, comparando la tabla sin la limpieza de variables con la tabla con la limpieza de variables, arrojando resultados de RMSE muy similares entre estas de 54.26 y 56.58 respectivamente, aumentado el error de las predicciones. Sin embargo al aplicar la métricas del R2, donde nos arrojó un mejor resultado en la tabla con la limpieza de variables, por lo que se decide seguir adelante con esta.

Luego se realizaron los modelos de Árboles de decisión y Random forest para comprobar diferentes resultados, siendo el Random forest con hiperparámetros ajustados el que mejor RMSE arroja, con un puntaje de 50.26 para los datos de entrenamiento y 53.54 para los datos de validación. Sin embargo es un resultado nos indica que el modelo aún no es confiable, por lo que se espera mejorar estos resultados.

Dificultades en el desarrollo

La base contenía varias variables que para efectos prácticos del proyecto no se les encontró utilidad o por el contrario dificultaron mucho el tratamiento de los datos, como la variable de código postal, que al ser categórica y contener tantos datos únicos nos creaba una base con más de 80.000 columnas a la hora de utilizar las variables dummies.