Buenas tardes, me gustaría presentarles mi trabajo práctico. El objetivo de mi trabajo es determinar si la presencia y las características de hoteles o negocios cercanos a una propiedad de alquiler turístico tiene un impacto en los precios de alquiler.

Para llevar a cabo mi trabajo, he aunado datos de alquiler turístico de corta duración (AirBnB) en la ciudad de Berlín con datos de Foursquare sobre la ubicación y característica de locales de interés en la misma ciudad con datos de hoteles obtenidos en TripAdvisor. Luego he comparado los precios de alquiler con la presencia o ausencia de hoteles o negocios cercanos. Además, también he considerado otros factores que podrían afectar los precios de alquiler, como la ubicación, el tamaño de la propiedad y las características de los locales (si son restaurantes, tiendas por ejemplo o si un hotel tiene 1 o 4 estrellas).

Para llevar a cabo mi trabajo, he utilizado un modelo de precios hedónicos para determinar el impacto de la presencia de hoteles o negocios cercanos y otros factores en los precios de alquiler turístico.

Inicialmente, he realizado una regresión lineal con las variables transformadas en logaritmos (el modelo tradicional de precios hedónicos), pero el R^2 obtenido no fue satisfactorio. Por lo tanto, decidí probar otros métodos de regresión. En primer lugar, utilicé una regresión con regularización ridge para ver si podía mejorar el ajuste del modelo. Sin embargo, aún no obtuve resultados satisfactorios. También lo hice con una regresión Lasso que tampoco me ha dado resultados significativos. Finalmente, utilicé un árbol de decisión gradient xboost para ajustar el modelo y, finalmente, obtuvimos un R^2 aceptable y resultados significativos. También he aplicado un árbol de decisión para ver si las variables seleccionadas eran determinantes en la determinación de precio (utilizando anova en vez de class). En este sentido, he trabajado con modelos de aprendizaje supervisado como propuesto en el enunciado de la práctica.

Uno de los principales desafíos que he enfrentado en mi trabajo ha sido la recopilación y transformación de datos. En primer lugar, intenté obtener información sobre hoteles a través de TripAdvisor utilizando Beautifulsoup, pero me encontré con el problema de que las latitudes y longitudes estaban vacías y no había direcciones disponibles para poder ubicar los hoteles en la página estática.

Para solucionar este problema, tuve que hacer uso de un web crawler para obtener las latitudes y longitudes de cada hotel. Luego, uní los precios, las estrellas y las reseñas obtenidas con Beautifulsoup con la latitud y la longitud obtenidas con el web crawler.

Además, también tuve que hacer uso de una API de Foursquare para obtener información sobre los locales cercanos a las propiedades de alquiler turístico. Me conecté al entorno de desarrollo de Foursquare y seleccioné las categorías que eran relevantes para mi análisis.

No obstante, la falta de la categoría de vecindario en TripAdvisor ha sido un obstáculo para poder conectar la información con Foursquare y así poder atribuir precios, estrellas y reseñas medias por vecindario.

Una posible estrategia que podrías haber seguido para abordar este problema es utilizar el algoritmo KNN (k-nearest neighbors) para asignar los vecindarios a los hoteles en TripAdvisor basándose en la cercanía de sus latitudes y longitudes con los vecindarios que tienen latitudes y longitudes en Foursquare. O mismo si tenían una determinada distancia, atribuirlo en un radius al vecindario definido en Foursquare. Sin embargo estas son técnicas supervisadas y quería hacer uso de una técnica no supervisada, que era parte del enunciado de la práctica.

Por ello he elegido utilizar el algoritmo k-means para atribuir los vecindarios a los hoteles en TripAdvisor y también para clasificar a Foursquare en base a sus latitudes y longitudes. De esta manera, pude hacer un merge de los dos conjuntos de datos utilizando esta nueva variable común. Con ello he podido conectar la información para luego aunar con la base de datos de Inside AirBnB que me da precios y características de apartamentos y casas para alquiler turístico y de ahí poder hacer el análisis de manera efectiva, a través de la conexión con latitud y longitud.

Se ha tenido que hacer uso del one-hot encoder para las variables dummies, agregar datos por vecindario y por tipo de locales entre otras transformaciones necesarias para llevar a cabo el análisis, que se describen en el código.

En resumen, he utilizado diferentes métodos de recogida de datos, k-means para atribución de vecindarios y poder aunar las bases de datos, y, luego análisis de regresión, y luego un árbol de decisión xboost, para determinar el impacto de la presencia de hoteles o negocios cercanos y otros factores en los precios de alquiler turístico. Los resultados obtenidos mediante estos métodos me han permitido llegar a conclusiones precisas y significativas sobre el tema de mi investigación. Esta información podría ser útil para dueños de propiedades de alquiler turístico que quieran maximizar su ingreso, para turistas que buscan encontrar el mejor precio posible para su alojamiento y para planificadores urbanos.