Proyecto Analítica de Datos

Entrega 1

Laura Victoria Ramos Agudelo

Tutor

Raúl Ramos Pollán, Professor of Computer Science

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería de sistemas

Medellín

Planteamiento del Problema

La crisis de la vivienda holandesa es uno de los mayores problemas a los que se enfrentan los residentes. Debido a múltiples factores, como el crecimiento de la población y la escasez de trabajadores de la construcción, la disponibilidad de viviendas ha disminuido significativamente. Esta disminución ha llevado el alquiler a precios altísimos, lo que hace que muchos se pregunten si se están aprovechando de ellos.

Para responder a esta pregunta, el modelo debe predecir el alquiler de una casa a partir de sus características (es decir, ubicación, tamaño, instalaciones, etc.). El propósito de estos datos es poder investigar tendencias y patrones en el mercado de alquiler de bienes raíces en los Países Bajos. Con suerte, estos datos pueden explicar las situaciones actuales y ayudar a comprender lo que sucederá en este mercado.

Dataset o Base de Datos

El Dataset seleccionado es de una competición de Kaggle llamada Netherlands Accommodation Prices (FCG) y puede consultar en el siguiente enlace: https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/data. Esta base de datos contiene toda la información disponible en https://kamernet.nl/ para cada propiedad. El sitio web fue rastreado diariamente y si aparecía una nueva propiedad, se añadía a la base de datos. Si se encontraba una propiedad que ya existía, se añadía a las fechas de publicación.

Este Dataset tiene más de 46,000 datos desde el 14 de julio de 2019 hasta el 3 de marzo de 2020 y se han dividido en dos grupos:

- Conjunto de entrenamiento (train.csv).
- Conjunto de prueba (test.csv).

El conjunto de entrenamiento debe usarse para construir el modelo de aprendizaje automático. Este archivo contiene 27.915 filas que incluyen el alquiler de cada alojamiento junto con otras 33 características:

Número de fila	Variables	Descripción
1	id	Identificador
2	title	Nombre del alojamiento
3	city	Ciudad

4	postalCode	Código postal
5	latitude	Latitud en grados
6	longitude	Longitud en grados
7	areaSqm	Tamaño en metros cuadrados
8	firstSeenAt	Hora de registro del propietario (AAAA-MM-DD HH-MM-SS)
9	lastSeenAt	Última aparición del propietario (AAAA-MM-DD
10	isRoomActive	Disponibilidad actual
11	rawAvailability	Periodo de tiempo de disponibilidad (DD-MM-AAAA)
12	postedAgo	Hace cuánto tiempo se publicó la propiedad
13	descriptionNonTranslated	Descripción original
14	descriptionTranslated	Descripción traducida
15	rentDetail	Justificación del alquiler
16	propertyType	Tipo de alojamiento
17	furnish	Presencia de muebles
18	energyLabel	Eficiencia energética
19	gender	Género del propietario
20	internet	Disponibilidad de internet
21	roommates	Número de compañeros de cuarto
22	shower	Disponibilidad de la ducha
23	toilet	Disponibilidad del baño
24	kitchen	Disponibilidad de la cocina
25	living	Disponibilidad de la sala de estar

26	pets	Política de mascotas
27	smokingInside	Política para fumadores
28	matchAge	Edad deseada del inquilino (rango en años)
29	matchGender	Género deseado del inquilino
30	matchCapacity	Número de personas que pueden vivir en el alojamiento
31	matchLanguages	Idioma deseado
32	matchStatus	Situación laboral/ocupación deseada
33	coverImageUrl	Url de la imagen de portada del alojamiento
34	rent	Valor del alquiler. Función objetivo

- El conjunto de datos de prueba (test.csv) debe usarse para ver cómo se desempeña el modelo en datos no vistos. Por lo tanto, no se proporciona el alquiler de cada alojamiento. El propósito del modelo es predecir estos valores. Este archivo tiene 18.610 filas.

A priori, consideramos eliminar del análisis las columnas que contienen vínculos a imágenes y texto descriptivo en formato plano, ya que esta información resulta complicada de manejar debido a nuestra limitada experiencia en este campo.

Métrica de Evaluación de Machine Learning

La métrica de evaluación de Machine Learning para el modelo será el Error Absoluto Promedio (MAE) el cual nos proporcionará el promedio de la diferencia absoluta entre la predicción del modelo y el valor objetivo. Esta métrica se calcula de la siguiente manera:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Donde:

yi = son la observaciones actuales de las series de tiempo

$\hat{y}_i$ = *es la seriede tiempo estimada o pronosticada*.

n = es el número de puntos de datos no faltantes

El objetivo general al trabajar con el MAE es minimizarlo. Un MAE bajo indica que el modelo está haciendo predicciones precisas y cercanas a los valores reales, lo que generalmente se considera un signo de un buen rendimiento del modelo.

Por otro lado, un Mean Absolute Error (MAE) alto en el contexto de un modelo de regresión o predicción significa que las predicciones generadas por el modelo tienden a estar bastante alejadas de los valores reales o de referencia. En otras palabras, un MAE alto indica que el modelo tiene dificultades para hacer predicciones precisas y que comete errores significativos en sus estimaciones.

Métrica de Evaluación del Negocio

Nuestro modelo de predicción podría servir para responder varias preguntas. ¿Cuál es la zona más cara para vivir?, ¿Qué tipo de propiedad ofrece mejor relación calidad-precio?, ¿Qué característica tiene el mayor impacto en el precio?. Igualmente, es útil para varios tipos de negocio, Si eres una empresa dedicada al alquiler de inmuebles, puedes usar este modelo para establecer el precio óptimo.Si eres un inquilino o una empresa dedicada a encontrar vivienda para quienes lo necesitan (como expatriados por ejemplo), deberías poder encontrar propiedades rentables buscando propiedades que se desvíen de este modelo.

En nuestro proyecto, la métrica clave para evaluar el éxito del negocio se enfoca en el aumento de los ingresos operacionales de la empresa que nos ha contratado, la cual se dedica al alquiler de propiedades. Nuestro modelo brinda la capacidad de determinar precios óptimos para las propiedades en alquiler.

Además, nuestro modelo ayuda a identificar propiedades que podrían estar subvaluadas o sobrevaloradas en el mercado. Esto permite a la empresa buscar oportunidades para adquirir propiedades a precios ventajosos o ajustar estratégicamente los precios de alquiler.

Desempeño Deseable en Producción

Según el departamento de marketing de FCG, un modelo de predicción del precio de alquiler de propiedades en los Países Bajos debería de tener un porcentaje de acierto de al menos 70%, ya que se usará el modelo para determinar precios óptimos para las propiedades en alquiler. Si el porcentaje de acierto es menor sería contraproducente porque los ingresos no compensarán los costos de tener el modelo en producción.

Cita	Netherlands Accommodation Prices (FCG)
Referencia Estilo IEEE (2020)	[1] Kaggle. (s. f.). https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview