Entrega N° 1 del Proyecto

Miguel Ángel Sánchez Peñates

Tutor

Raul Ramos Pollan, Professor of Computer Science

Introducción a la Inteligencia Artificial para las Ciencias e Ingeniería

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería Industrial

Medellín, Antioquia, Colombia

2023

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, debe predecir el alquiler de una casa a partir de sus datos (es decir, ubicación, tamaño, instalaciones, etc.).

Dataset o Base de Datos

El dataset seleccionado es de una competición de Kaggle llamada **Precios de Alojamientos en Países Bajos,** la cual podemos consultar en el siguiente enlace:

https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview.

Los datos de esta competición 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. Para el conjunto de entrenamiento, proporcionamos el alquiler de cada alojamiento junto con otras 35 características.

El conjunto de prueba 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.

Para ilustrar el formato de un archivo de envío, proporcionamos **sample_submission.csv**

La carpeta contiene los siguientes archivos

- **train.csv** - el conjunto de entrenamiento
- **test.csv** - el conjunto de prueba
- **sample_submission.csv**: un archivo de envío de muestra en el formato correcto

El archivo contiene las siguientes variables:

Variables	**Descripción**

Titulo	Nombre del alojamiento
Ciudad	Nombre de la ciudad
Código postal	Código postal
Latitud	Latitud en grados
Longitud	Longitud en grados
Área m^2	Tamaño en metros cuadrados
Visto por primera vez	Hora de registro del titular (AAAA-MM-DD HH-MM-SS) GMT
Visto por última vez	Última aparición del propietario (AAAA-MM-DD HH-MM-SS) GMT
isRoomActive	Disponibilidad actual
rawDisponibilidad	Periodo de tiempo de disponibilidad (DD-MM-AAAA)
Publicado hace	Hace cuánto tiempo se planteó la propiedad
Descripción no traducida	Descripción original
Descripción Traducido	Descripción traducida
Limpiar Detalle	Justificación del alquiler
Tipo de propiedad	Tipo de alojamiento
Amoblar	Presencia de muebles
Etiqueta de energía	Eficiencia de energética
Genero	Sexo del propietario
Internet	Disponibilidad de internet
Compañeros de cuarto	Numero de compañeros de cuarto
Ducha	Propiedad de la ducha
Baño	Propiedad del baño
Cocina	Propiedad de la cocina
Viviendo	Propiedad de la sala de estar
Mascotas	Mascotas permitidas
Fumar	Fumar está permitido
Edad	Edad permitida del inquilino
Coincidencia de genero	Sexo del inquilino deseado

Capacidad de coincidencia	# de personas que pueden vivir en el alojamiento
Coincidir idioma	Idioma deseado
Estado	Estado deseado
coverImageUrl	Url de la imagen de portada del alojamiento
Alquilar	Función objetivo

Métrica de evaluación

La métrica de evaluación 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 metrica se calcula de la siguiente manera:

$$MAE = \frac{(\sum_{i=1}^{n} |y_i - \overline{y_i}|)}{n}$$

Donde:

$y_i = son\ la\ observaciones\ actuales\ de\ las\ series\ de\ tiempo.$

$\overline{y_i} = es\ la\ seriede\ tiempo\ estimada\ o\ pronosticada.$

$n = es\ el\ número\ de\ puntos\ de\ datos\ no\ fal\ \text{tan}\ t\ es$

Es importante resaltar que el MAE tiene un umbral predeterminado con un límite superior del 80%, concluyéndose que:

- Tendencia al alza: Una tendencia al alza indica que la métrica se está deteriorando. Los datos de comentarios ya son significativamente distintos respecto a los datos de entrenamiento.
- Tendencia a la baja: Una tendencia a la baja indica que la métrica está mejorando. Esto significa que el reentrenamiento del modelo es efectivo.
- Variación errática o irregular: Una variación errática o irregular indica que los datos de comentarios no son coherentes entre evaluaciones. Incremente el tamaño mínimo de la muestra para el supervisor de calidad.

Desempeño

Lo que se espera pronosticar es que el precio de los alojamientos respectos a los datos de entrenamiento es que exista una tendencia a la baja, lo que nos representaría que el modelo es adecuado para determinar si los precios de los alquileres están ajustados a las características de cada alojamiento. Es decir, si el precio de un alquiler es 1000€ y el modelo predice 1200€, entonces el error es del 120%. Pero si el precio es de 400€ y el modelo predice 200€ el error es del 50%. Como métrica de negocio se podría usar el incremento en ventas gracias a la utilización del modelo.

Bibliografia

Netherlands Accommodation Prices (FCG) | Kaggle. (s. f.). https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview/evaluation