Entrega N° 2 del Proyecto

Miguel Ángel Sánchez Peñates
Daniel Góngora García
Jhon Alexander Longas

Tutor
Raúl 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.

Análisis Exploratorio de los datos

Antes de entrenar un modelo predictivo, o incluso antes de realizar cualquier cálculo con un nuevo conjunto de datos, es muy importante realizar una exploración descriptiva de los mismos. Este proceso permite entender mejor qué información contiene cada variable, así como detectar posibles errores. Algunos ejemplos frecuentes son:

- Que una columna se haya almacenado con el tipo incorrecto: una variable numérica está siendo reconocida como texto o viceversa.
- Que una variable contenga valores que no tienen sentido: por ejemplo, para indicar que no se dispone del precio de un alojamiento se introduce el valor 0 o un espacio en blanco.
- Que en una variable de tipo numérico se haya introducido una palabra en lugar de un número.

Además, este análisis inicial puede dar pistas sobre qué variables son adecuadas como predictores en un modelo.

Cargamos la base de datos

id	title	city	postalCode	latitude	longitude	areaSqm	firstSeenAt	lastSeenAt	isRoomActive	rawAvailability	...	living
0	West-Varkenoordseweg	Rotterdam	3074HN	51.896601	4.514993	14	2019-07-14 11:25:46.511000+00:00	2019-07-26 22:18:23.142000+00:00	True	26-06-'19 - Indefinite period	...	None
3	Ruiterakker	Assen	9407BG	53.013494	6.561012	16	2019-07-14 11:25:46.988000+00:00	2019-07-18 22:00:31.174000+00:00	False	16-06-'19 - Indefinite period	...	None
8	Brusselseweg	Maastricht	6217GX	50.860841	5.671673	16	2019-07-14 11:25:47.814000+00:00	2019-08-10 00:14:27.130000+00:00	True	15-07-'19 - Indefinite period	...	None
10	Donkerslootstraat	Rotterdam	3074WL	51.893195	4.516478	25	2019-07-14 11:25:48.140000+00:00	2019-07-16 06:05:32.183000+00:00	False	01-08-'19 - Indefinite period	...	None
12	Vorselenburgstraat	Alphen aan den Rijn	2405XJ	52.122335	4.661434	10	2019-07-14 11:25:48.465000+00:00	2019-08-01 00:02:40.516000+00:00	True	08-07-'19 - Indefinite period	...	None

5 rows × 33 columns

Se realizo una descripción del tipo de variable

```
title                    object
city                     object
postalCode               object
latitude                 float64
longitude                float64
areaSqm                  int64
firstSeenAt              object
lastSeenAt               object
isRoomActive             object
rawAvailability          object
postedAgo                object
descriptionNonTranslated object
descriptionTranslated    object
rentDetail               object
propertyType             object
furnish                  object
energyLabel              object
gender                   object
internet                 object
roommates                object
shower                   object
toilet                   object
kitchen                  object
living                   object
pets                     object
smokingInside            object
matchAge                 object
matchGender              object
matchCapacity            object
matchLanguages           object
matchStatus              object
coverImageUrl            object
rent                     int64
dtype: object
```

Número de observaciones y valores ausentes

Junto con el estudio del tipo de variables, es básico conocer el número de observaciones disponibles y si todas ellas están completas. Los valores ausentes son muy importantes a la hora de crear modelos, la mayoría de los algoritmos no aceptan observaciones incompletas o bien se ven muy influenciados por ellas. Aunque la imputación de valores ausentes es parte del preprocesado y, por lo tanto, debe de aprenderse únicamente con los datos de entrenamiento, su identificación se tiene que realizar antes de separar los datos para asegurar que se establecen todas las estrategias de imputación necesarias.

- Dimensión de la base de datos: (27915, 33).

- Numero de datos ausentes por variables:

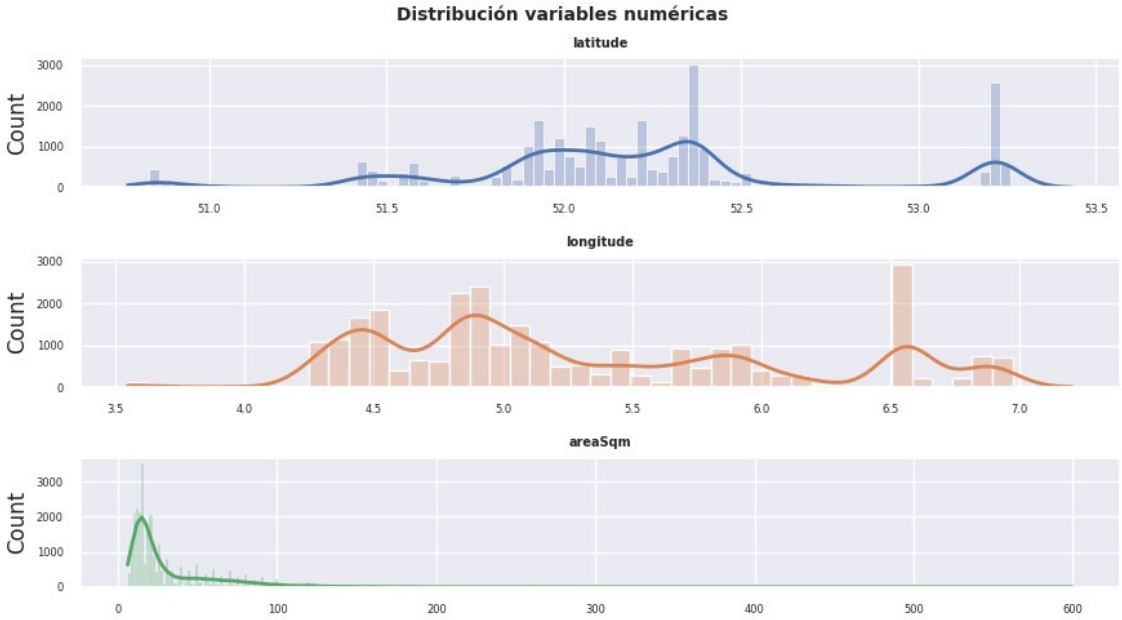
```
title                        0
coverImageUrl                0
propertyType                 0
rawAvailability              0
lastSeenAt                   0
firstSeenAt                  0
rent                         0
longitude                    0
latitude                     0
postalCode                   0
city                         0
areaSqm                      0
postedAgo                    6
matchStatus                 63
matchLanguages              63
matchCapacity               63
matchGender                 63
matchAge                    63
smokingInside               63
pets                        63
living                      63
kitchen                     63
energyLabel                 63
shower                      63
internet                    63
isRoomActive                63
toilet                      63
descriptionNonTranslated   111
furnish                    214
gender                     536
roommates                  536
rentDetail                7896
descriptionTranslated    10140
dtype: int64
```

Análisis de la variable respuesta

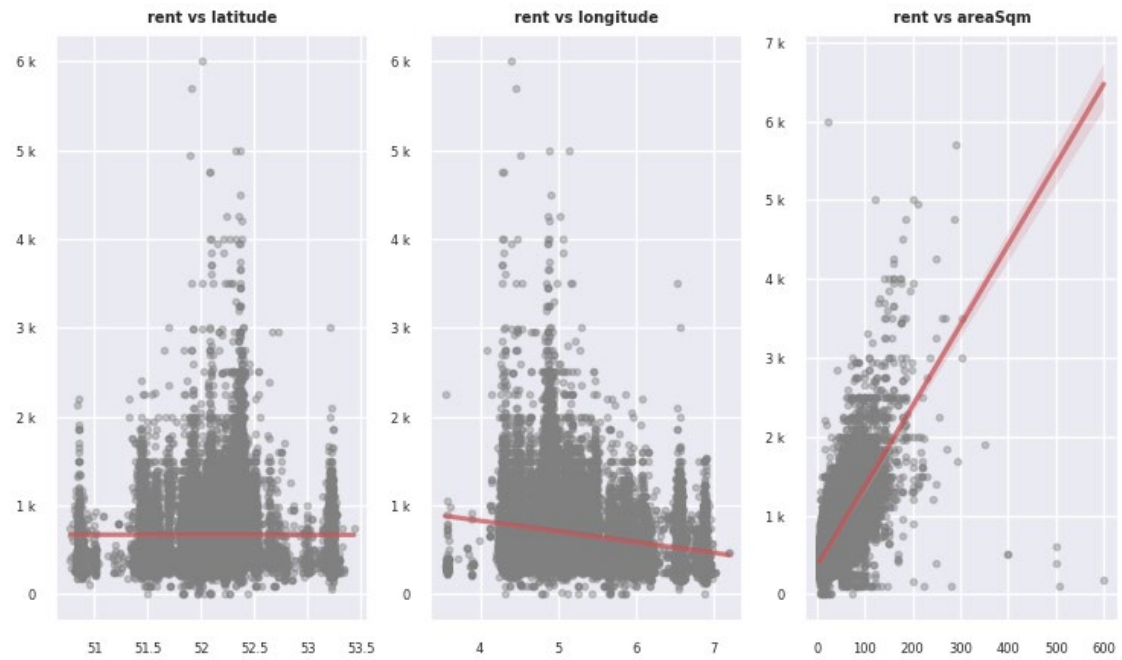
Cuando se crea un modelo, es muy importante estudiar la distribución de la variable respuesta, ya que, a fin de cuentas, es lo que interesa predecir. La variable alquiler (rent) tiene una distribución asimétrica con una cola positiva debido a que, unos pocos alojamientos, tienen un precio muy superior a la media. Este tipo de distribución suele visualizarse mejor si se calcular la asimetría de los datos, que en este caso es 2.36.

Descripción de las variables numéricas

Se realizo los gráficos de la distribución de cada una de las variables numéricas



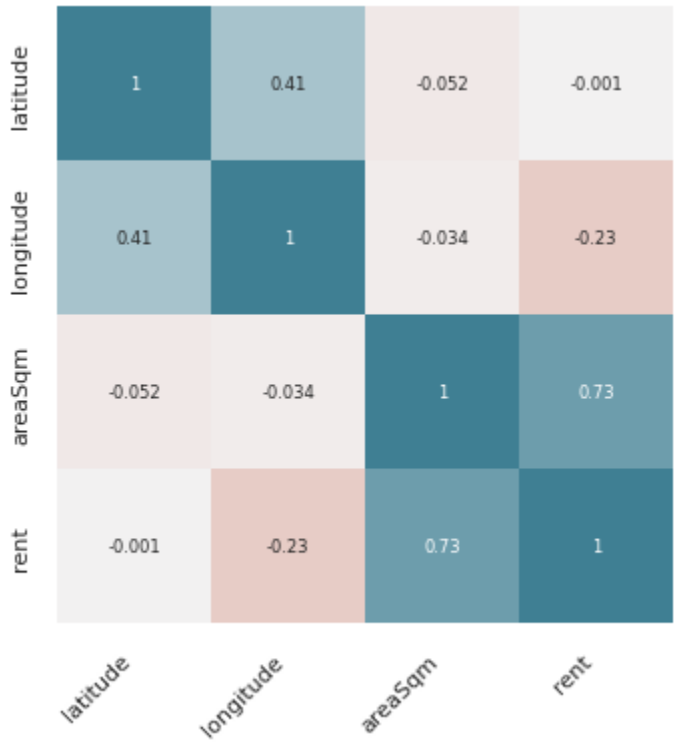
Como el objetivo del estudio es predecir el precio del alquiler de los alojamientos, el análisis de cada variable se hace también en relación con la variable respuesta precio. Analizando los datos de esta forma, se pueden empezar a extraer ideas sobre qué variables están más relacionadas con el precio y de qué forma.



Correlación de las variables numéricas

Algunos modelos (LM, GLM, ...) se ven perjudicados si incorporan predictores altamente correlacionados. Por esta razón, es conveniente estudiar el grado de correlación entre las variables disponibles.

	variable_1	variable_2	r	abs_r
11	areaSqm	rent	0.729018	0.729018
14	rent	areaSqm	0.729018	0.729018
1	latitude	longitude	0.405798	0.405798
4	longitude	latitude	0.405798	0.405798
7	longitude	rent	-0.230255	0.230255
13	rent	longitude	-0.230255	0.230255
2	latitude	areaSqm	-0.052129	0.052129
8	areaSqm	latitude	-0.052129	0.052129
6	longitude	areaSqm	-0.033980	0.033980
9	areaSqm	longitude	-0.033980	0.033980



Primeros Modelos

Como primer modelo se ajustó una regresión logística, obteniendo resultados desfavorables, sin embargo, antes de entrenar el modelo con los datos, se seleccionaron los features a utilizar por concenso entre los integrantes del equipo:

```
LAS VARIABLES SELECCIONADAS PARA INCLUIR EN EL PRIMER ACERCAMIENTO A UN MODELO PREDICTIVO, SON:

areaSqm
rentDetail
propertyType
furnish
gender
internet
shower
toilet
kitchen
living
pets
smokingInside

Y como variable de respuesta rent, vamos a clasificar cada valor por el cuartil al que pertenece
```

También se definió la función *data_prep()* para crear X_train e y_train con los features seleccionados.

Finalmente se ajusto el modelo de regresión logística, obteniendo el siguiente resultado:

```
Accuracy for train=  56.12400275547991
```

Como trabajo futuro se van a explorar mas modelos de machine learning que se ajusten mejor a los datos, además de realizar una selección de variables.

Bibliografia

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