



Instituto Tecnológico
de Buenos Aires


AIRBNB RATING

Agustin Lara

—

Analítica Predictiva

12/06/2023

12	—	AGUSTIN LARA ACOSTA		0.66890	3	6d
----	---	---------------------	---	---------	---	----



AGENDA

01 Análisis del Dataset

02 Columnas agregadas

03 Modelos testeados

04 Conclusiones

01

Análisis del Dataset

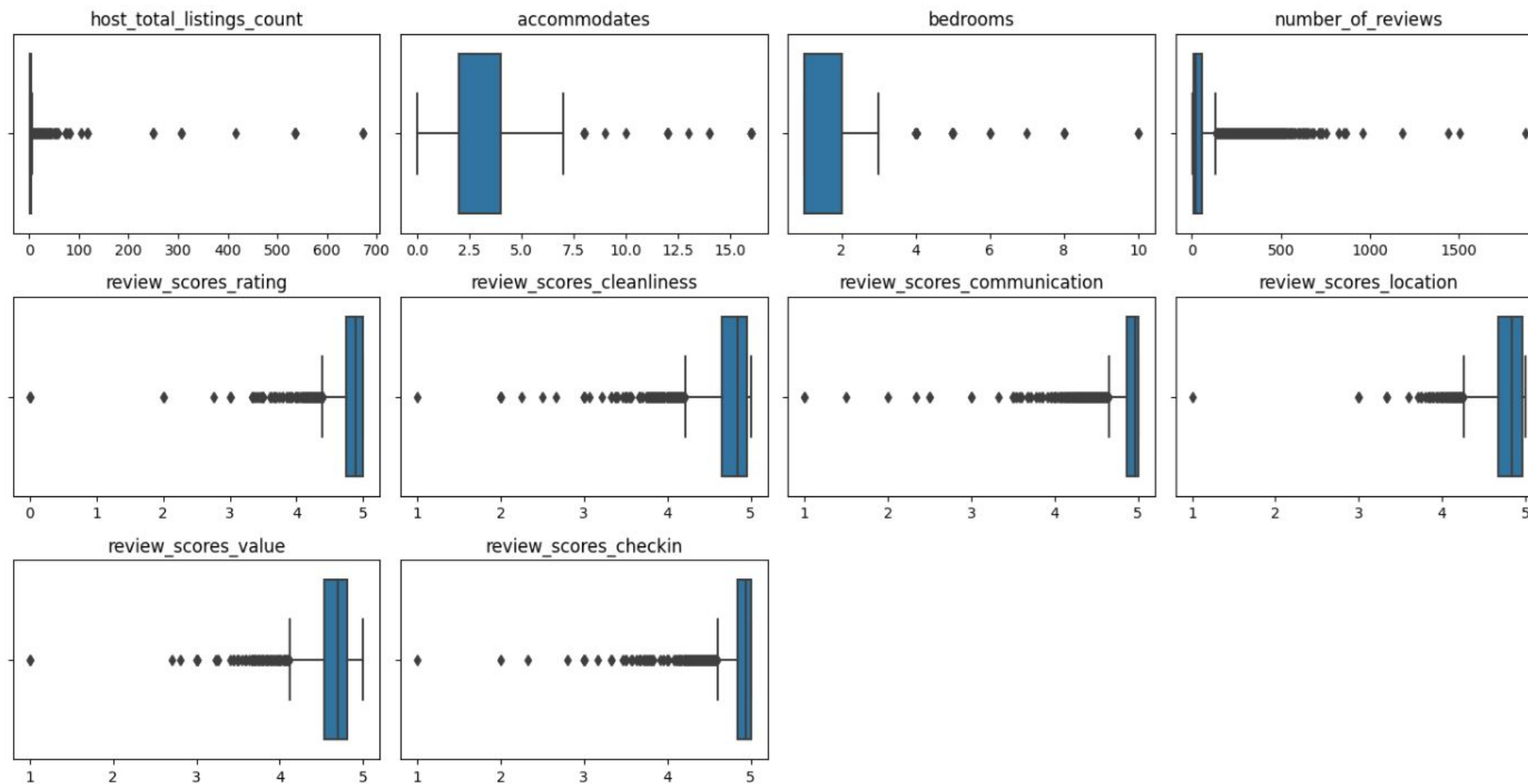
- Tanto la columna *Bedroom* y *Beds* presentaban valores nulos
 - Sustitución por el valor entero del promedio de los valores
- *host_is_superhost* / *host_has_profile_pic* / *host_identity_verified* / *has_availability* / *instant_bookable*
 - Se reemplazaron *t* por 1 y *f* por 0
- *host_since* / *last_review*
 - Se convirtieron a datetime
- *host_response_rate* / *host_acceptance_rate* / *price*
 - Se eliminó '%' y se convirtió a float
 - Se eliminó '\$', '.00' y se convirtió a int
- *neighborhood_overview* / *amenities*
 - se cambiaron las regex que contenían '
' y '', '\u2013' y '\u2019'

01

Análisis del Dataset

- *'source', 'neighbourhood_zone', 'host_response_time', 'host_verifications', 'neighbourhood_cleansed', 'property_type', 'bathrooms_text', 'room_type'*
 - Fueron codificadas

01 Análisis del Dataset



02 Columnas agregadas

Aquí tienes las etiquetas de ubicación geográfica correspondientes a los barrios de Ámsterdam que mencionaste:

'Centrum-Oost': centro, este

'Centrum-West': centro, oeste

'Oud-Oost': antiguo este

'De Pijp - Rivierenbuurt': sur

'Noord-Oost': norte, este

'Oud-Noord': antiguo norte

'Noord-West': norte, oeste

'De Aker - Nieuw Sloten': oeste

'Geuzenveld - Slotermeer': oeste

'Bijlmer-Centrum': sureste

'Buitenveldert - Zuidas': sur

'Westerpark': oeste

'Slotervaart': oeste

'De Baarsjes - Oud-West': oeste

'Bos en Lommer': oeste

← → ↺ 🔒 location.foursquare.com/places/docs/categories

Welcome ▾

Places Data ▴

Places Data Overview

How Does Places Work?

▾ Places Data Types

Places Attributes

Categories

Chains

Supported Countries

Places Delivery ▾

Category Taxonomy

Foursquare Places includes a hierarchical taxonomy of categories from which each POI record is classified. The ten parent categories are:

• Arts and Entertainment

• Business and Professional Services

• Community and Government

• Dining and Drinking

• Event

• Health and Medicine

• Landmarks and Outdoors

• Retail

• Sports and Recreation

• Travel and Transportation

Aquí tienes la tabla completa con las etiquetas correspondientes en la columna "Clasificación":

Elemento	Apariciones	Clasificación
Essentials	4715	Essentials
Smoke alarm	4421	Well Being
Wifi	4337	Administration
Heating	4101	Heating/Cooling
Hot water	3990	Essentials
Hangers	3949	Essentials
Hair dryer	3913	Bath & Cleaning
Kitchen	3841	Kitchen
Dishes and silverware	3781	Kitchen
Refrigerator	3630	Kitchen
Iron	3618	Essentials
Long term stays allowed	3583	Administration
Shampoo	3412	Bath & Cleaning
Bed linens	3215	Essentials
Cooking basics	3060	Kitchen
Carbon monoxide alarm	2810	Well Being
Coffee maker	2715	Kitchen
Fire extinguisher	2628	Well Being
Dishwasher	2435	Kitchen
Washer	2432	Bath & Cleaning
Private entrance	2407	Accessibility
First aid kit	2376	Well Being
Oven	2365	Kitchen
Microwave	2319	Kitchen
Stove	2100	Kitchen
Dedicated workspace	1813	Miscellaneous
Extra pillows and blankets	1796	Essentials
TV	1616	Entertainment
Host greets you	1594	Administration
Luggage dropoff allowed	1581	Administration
Cleaning products	1508	Bath & Cleaning
Shower gel		Bath & Cleaning
Dryer		Bath & Cleaning

Regenerate response

Send a message.

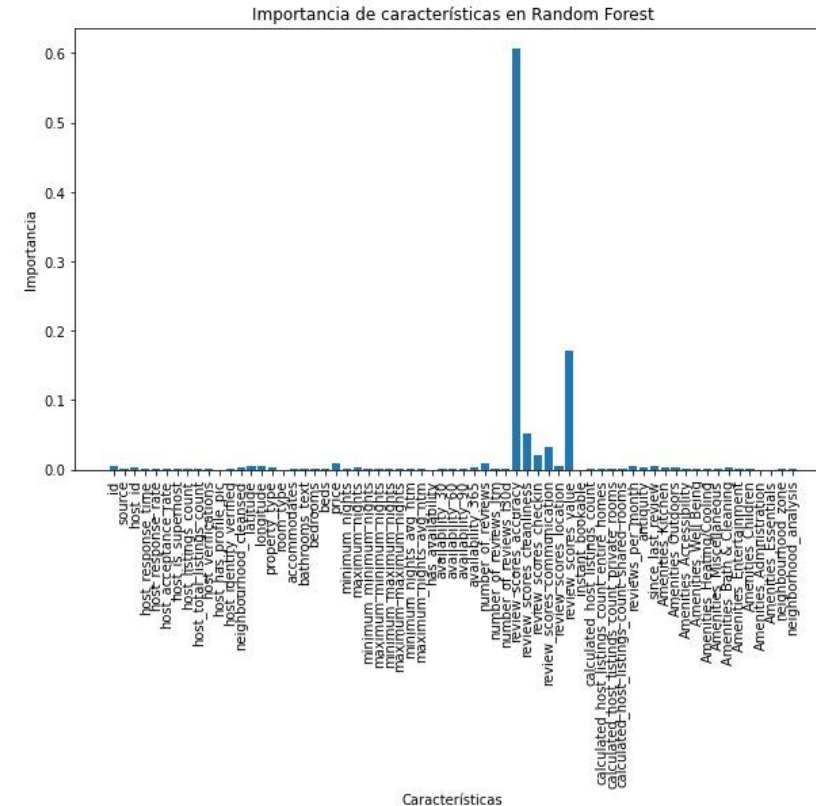
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT May 24 Version](#)

02 Columnas agregadas

- Se creó la columna *neighbourhood_zone* con las zonas que devolvió Chat GPT
- Se realizó un análisis con NLTK para *neighborhood_overview*, se buscó definir si el comentario era: muy positivo, positivo, neutro, negativo, muy negativo
- Se crearon las columnas *antiquity* y *since_last_review*
 - Ambas son la diferencia entre hoy, y la fecha de alta o de la última review (en días)
- Se crearon 11 columnas nuevas, *Amenities_cateogría*
 - A partir de la clasificación de Chat GPT, muestra la cantidad de amenities de esa categoría que posee el inmueble
- Se intentó generar una columna de POI, sin éxito por falta de API Key



- Buenos resultados, pero difieren en Kaggle.
MSE muy bajo
¿Podría tener **overfitting**?



03 Modelos testeados

- Random Forest
 - Limpieza de parámetros -> sin resultados
- CatBoost
 - Se cambiaron los grid search incorporando variaciones en el random_strength y bagging temperature -> resultados muy similares



CONCLUSIONES

- Pareciera que el overfitting no fue problema
- Muchas de las columnas agregadas no aportan información, es probable que estén generando ruido innecesario