Problem Set #2
Feature Engineering y Modelos Lineales

Melanie Álvarez Baldeón

airbnb

Compañía que ofrece una plataforma digital dedicada a la oferta de alojamientos a particulares y turísticos mediante la cual los anfitriones pueden publicitar y contratar el arriendo de sus propiedades con sus huéspedes.

Análisis Exploratorio Inicial

El problema principal fueron los valores nulos, varias categorías presentaron este inconveniente. No se encontraron duplicados.

#	Column	Non-Null Count	Dtype
0	id	74111 non-null	int64
1	log_price	74111 non-null	float64
2	property_type	74111 non-null	object
3	room_type	74111 non-null	object
4	amenities	74111 non-null	object
5	accommodates	74111 non-null	int64
6	bathrooms	73911 non-null	float64
7	bed_type	74111 non-null	object
8	cancellation_policy	74111 non-null	object
9	cleaning_fee	74111 non-null	bool
10	city	74111 non-null	object
11	description	74111 non-null	object
12	first_review	58247 non-null	object
13	host_has_profile_pic	73923 non-null	object
14	host_identity_verified	73923 non-null	object
15	host_response_rate	55812 non-null	object
16	host_since	73923 non-null	object
17	instant_bookable	74111 non-null	object
18	last_review	58284 non-null	object
19	latitude	74111 non-null	float64
...			
27	bedrooms	74020 non-null	float64
28	beds	73980 non-null	float64

Data Wrangling

API Geocoding

Uso del API de Google Maps para completar aquellos zipcodes NaN, esto con los datos de longitud y latitud.

Bedrooms NaN o 0

Según el EDA, en general, el número de bedrooms era la mitad del número de accommodates, por lo que se aplicó este reemplazo.

review_scores_rating

Se reemplazaron los valores nulos con la media de los ratings

Outliers en log_price

Usando el rango intercuartil, se localizaron outliers en la variable log_price.
El rango intercuartil fue el siguiente:

lower_bound	upper_bound
2.9631865462232887	6.574657392391346

El mayor valor de los outliers fue:

```
7.6004023345004
```

Puesto que la diferencia no es muy significativa, no se eliminaron los outliers.

Feature Engineering

Se escogió el siguiente conjunto de características:

#	Column	Non-Null Count	Dtype
0	id	74111 non-null	int64
1	log_price	74111 non-null	float64
2	property_type	74111 non-null	object
3	room_type	74111 non-null	object
4	amenities	74111 non-null	object
5	accommodates	74111 non-null	int64
6	cleaning_fee	74111 non-null	bool
7	city	74111 non-null	object
8	review_scores_rating	74111 non-null	float64
9	bedrooms	74111 non-null	float64

One-hot Encoding room_type

room_type contiene tres categorías, por lo que se crearon columnas para cada tipo

Label Encoding property_type

Se encontraron 17 tipos de propiedad, por lo que one-hot encoding no era la mejor opción.

True/false por 0/1

Los valores booleanos de cleaning fee fueron reemplazados por 0 y 1.

Amenities

Si la lista de amenities estaba vacía, se colocaba un 0, caso contrario, un 1.

Nota: no se consideró a zipcode porque el número no se refiere a un valor entero sino a una localidad. Esto podía crear una falsa relación de orden en los modelos.

Model Training and Evaluation

Después de la validación cruzada y probar varios modelos, se obtuvo que la regresión polinomial de segundo orden era el mejor modelo de predicción.

Model	RMSE_CV	MAE_CV	R2_CV	RMSE_Test
Polynomial Regression	0.479303	0.360298	0.553689	0.477713
Linear Regression SVD	0.489320	0.368183	0.534853	0.488243
Ridge Regression	0.489320	0.368183	0.534853	0.488243
Linear Regression	0.489323	0.368200	0.534849	0.488297
Batch Gradient Descent	0.520300	0.391022	0.473843	0.503889
Stochastic Gradient Descent	0.519211	0.389761	0.476330	0.509075
Lasso Regression	0.717548	0.561747	−0.000148	0.716788

1. RMSE_CV (0.4793)
Comparado con los otros modelos, Polynomial Regression tiene el RMSE_CV más bajo, lo que indica que en validación cruzada hizo mejores predicciones.

2. MAE_CV (0.3603)
Polynomial Regression hace predicciones más precisas en promedio.

3. R^2_CV (0.5537)
El modelo explica el 55.37% de la variabilidad en los datos.

4. RMSE_Test (0.4777)
Polynomial Regression tiene el mejor desempeño en datos no vistos.

Conclusiones

Si bien Polynomial Regression demostró ser el mejor modelo, no describe en su totalidad la variabilidad de los datos.
Sería útil probar a entrenar el modelo con más características, cuidando siempre el sobreajuste.

Para optimizar precios en la plataforma, Airbnb podría usar modelos más avanzados como árboles de decisión o redes neuronales, ya que los modelos lineales pueden no capturar toda la complejidad del mercado.

Los precios de Airbnb pueden tener cierta complejidad no capturada por modelos lineales, la incorporación de características polinómicas ayuda a mejorar la precisión de la predicción.