



MACHINE LEARNING

Proyecto Machine Learning

THE BRIDGE



AGUSTÍN GEROME
Data Science



PREDICCIÓN DEL PRECIO DE UN INMUEBLE

Elaboración de un **Modelo de Regresión** para analizar el precio de venta de un inmueble.

Objetivos del proyecto:

- Realizar búsquedas automatizadas de oportunidades inmobiliarias
- Poder realizar cotizaciones online del inmueble

CONTENIDO



1

Extracción
de datos

2

Análisis y
preparación

3

Elección del
Modelo

4

Conclusiones y
Próximos pasos

1 EXTRACCIÓN DE DATOS

API
idealista

EXTRACCIÓN DATOS

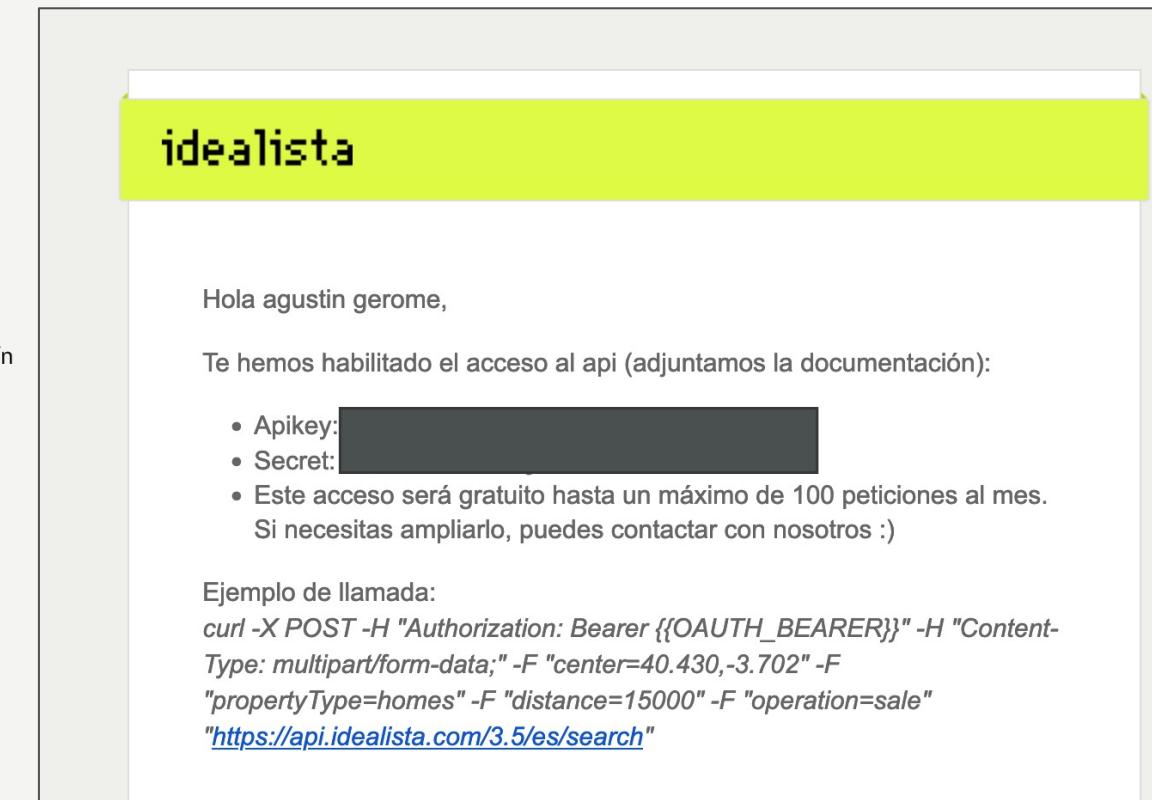
Contenidos

- ✓ Código identificador
- ✓ Localización
- ✓ Tipo de operación: venta o alquiler
- ✓ Fecha de creación y desactivación
- ✓ Precio total y unitario
- ✓ Tipología y subtipología
- ✓ Superficie útil y construida
- ✓ Estado de conservación
- ✓ Número de habitaciones y baños
- ✓ Disponibilidad de parking, trastero, piscina, ascensor, jardín
- ✓ Certificado energético
- ✓ Instalaciones de aire acondicionado y calefacción
- ✓ Orientación
- ✓ Número de visitas recibidas
- ✓ Número de contactos: enviados a un amigo, favoritos, contactos por email
- ✓ Información catastral
- ✓ Distancia al origen seleccionado

idealista

API de Testigos

<https://www.idealista.com/data/productos/desarrollo/api-de-testigos>



Parámetros:

- Tipo de operación
 - Tipo de propiedad
 - Coordenadas del centro
 - Máxima distancia
 - Precio Máximo
- Etc.

2 ANÁLISIS Y PREPARACIÓN

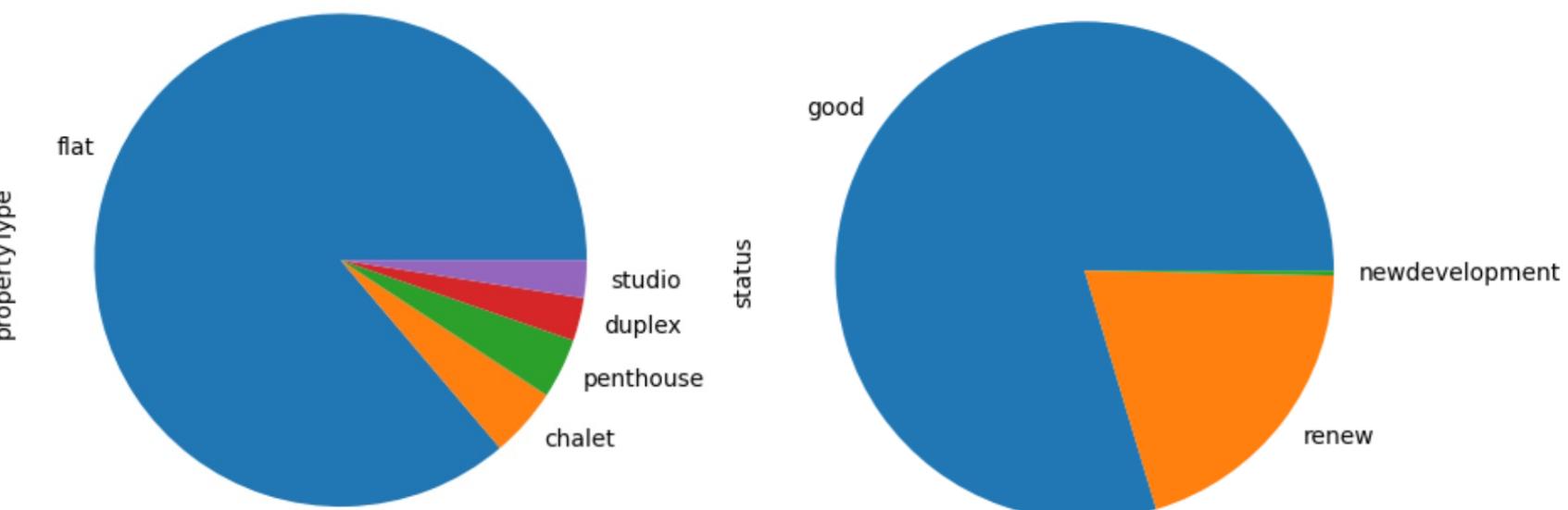
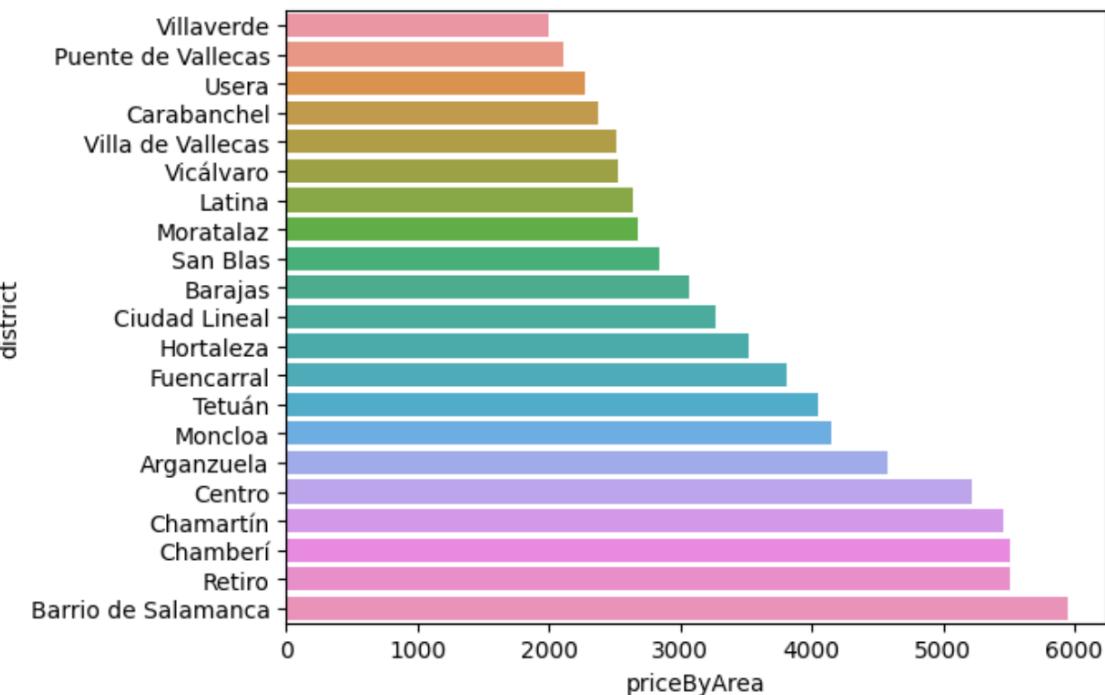
TIPO DE OPERACIÓN: VENTA
COORDENADAS: '40.4167,-3.70325' - MADRID
PRECIO MÁXIMO: 1.000.000 €

ANÁLISIS Y PREPARACIÓN

Int64Index: 1950 entries, 4156 to 1863

Data columns (total 40 columns):

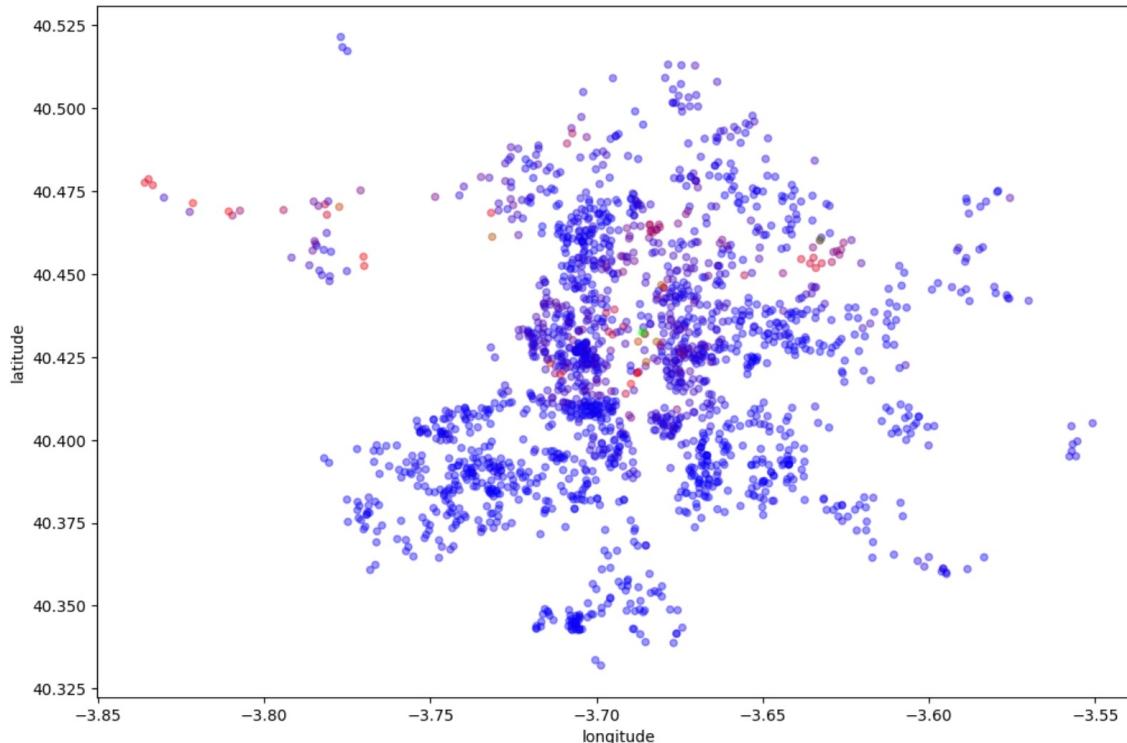
#	Column	Non-Null Count	Dtype
0	Unnamed: 0	1950	non-null
1	propertyCode	1950	non-null
2	thumbnail	1943	non-null
3	externalReference	1462	non-null
4	numPhotos	1950	non-null
5	propertyType	1950	non-null
6	operation	1950	non-null
7	size	1950	non-null
8	exterior	1950	non-null
9	rooms	1950	non-null
10	bathrooms	1950	non-null
11	address	1950	non-null
12	province	1950	non-null
13	municipality	1950	non-null
14	country	1950	non-null
15	latitude	1950	non-null
16	longitude	1950	non-null
17	showAddress	1950	non-null
18	url	1950	non-null
19	distance	1950	non-null
20	description	1947	non-null
21	hasVideo	1950	non-null
22	status	1950	non-null
23	newDevelopment	1950	non-null
24	parkingSpace	496	non-null
25	priceByArea	1950	non-null
26	detailedType	1950	non-null
27	suggestedTexts	1950	non-null
28	hasPlan	1950	non-null
29	has3DTour	1950	non-null
30	has360	1950	non-null
31	hasStaging	1950	non-null
32	topNewDevelopment	1950	non-null
33	superTopHighlight	1950	non-null
34	floor	1764	non-null
35	district	1950	non-null
36	neighborhood	1950	non-null
37	hasLift	1830	non-null
38	labels	190	non-null
39	newDevelopmentFinished	6	non-null



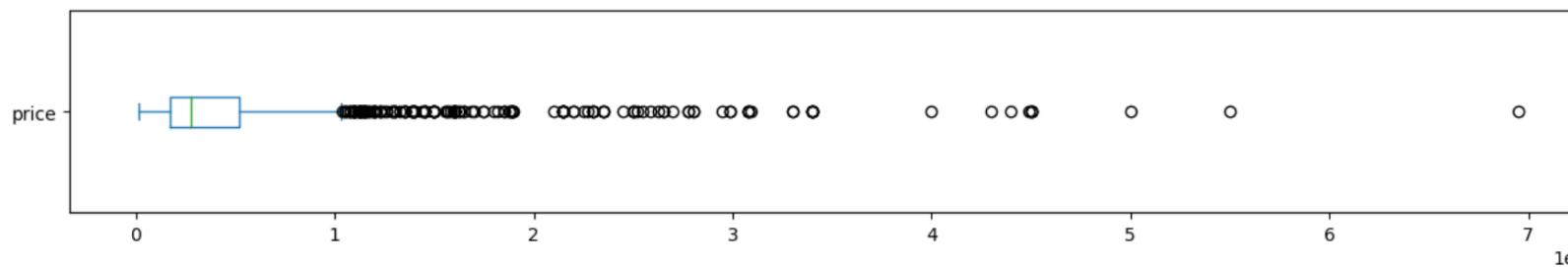
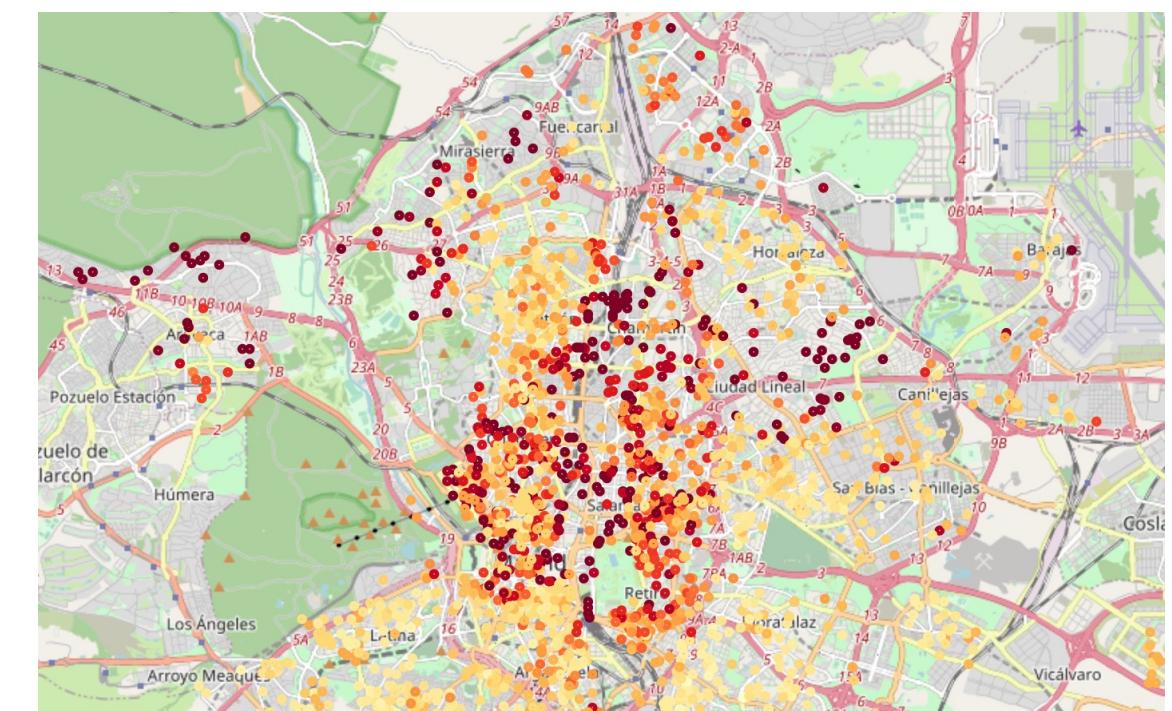
ANÁLISIS Y PREPARACIÓN

Distribución Geográfica limitada a MADRID

Con Outliers



Sacando Outliers (max 1.000.000€)



size	district	price	
1062	742.0	Barrio de Salamanca	6950000.0

ANÁLISIS Y PREPARACIÓN

ELIMINAR COLUMNAS

Columnas con los parámetros de la búsqueda especificada

- 'operation': Especifiqué en la búsqueda que es todo sale
- 'country': Todos son España
- 'municipality', 'province': es siempre Madrid

Columnas con info interna de Idealista que no me sirven para realizar el análisis y no influyen en el precio

- 'Unnamed: 0', 'propertyCode', 'thumbnail', 'externalReference', 'numPhotos','url','hasPlan', 'has3DTour', 'has360', 'hasVideo','hasStaging'

Columnas estrechamente relacionadas

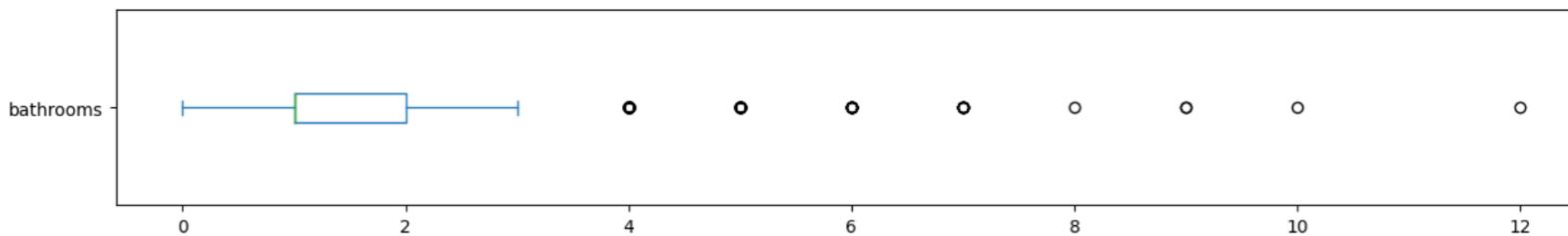
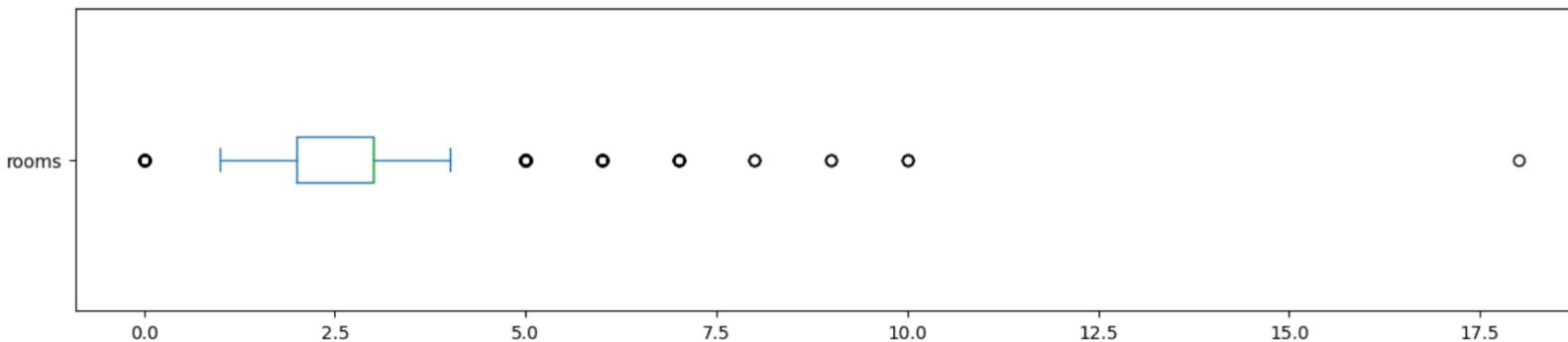
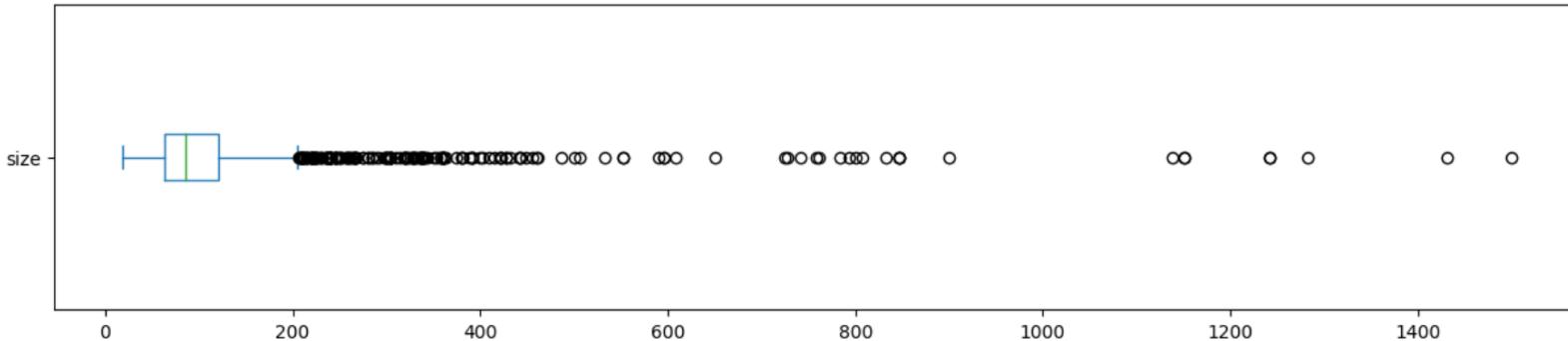
- 'address', 'showaddress', 'distance': Me voy a guiar por latitud y longitud
- 'priceByArea': estrechamente correlacionada con label ('price'/'size')
- 'description' : info duplicada. La info relevante de esta columna se encuentra en 'detailedType'
- 'suggestedTexts' : info duplicada. La info relevante de esta columna se encuentra en 'detailedType'

Columnas con demasiados missings/False o que no me proporciona información relevante

- 'labels', 'newDevelopmentFinished': Demasiados missings
- 'topNewDevelopment' : tiene 100% False
- 'superTopHighlight' : tiene 100% False
- 'newDevelopment': tiene 99,5% de False

ANÁLISIS Y PREPARACIÓN

OUTLIERS



3

ELECCIÓN DEL MODELO

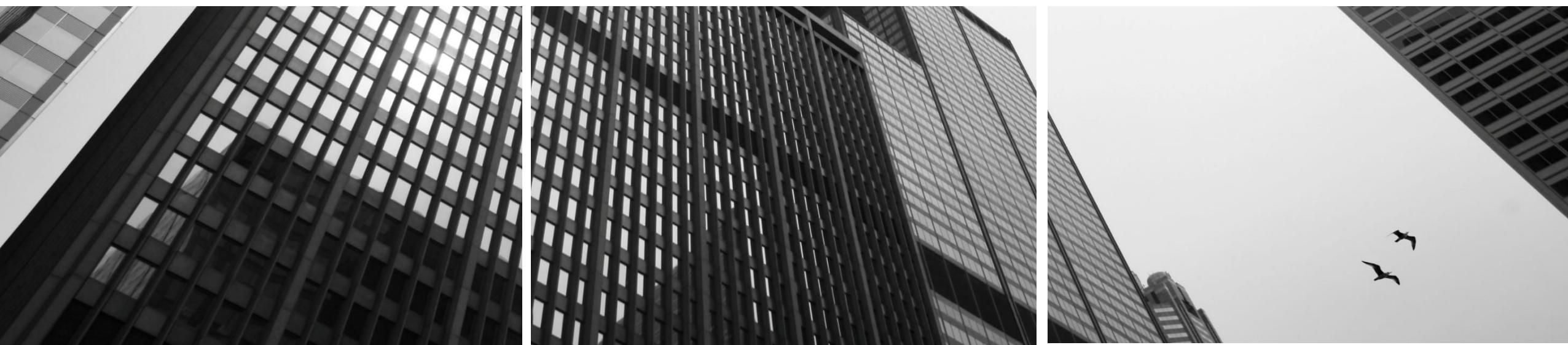


CatBoost

ELECCIÓN DEL MODELO

Comparación de modelos

	MODELOS TRAIN	Mean_Absolute_error	Root_Mean_Squared_error		MODELOS TEST	Mean_Absolute_error	Root_Mean_Squared_error
0	RandomForestRegressor	97169.423838	178183.256629	0	RandomForestRegressor	104680.510156	201510.346343
1	XGboost	79678.012067	139384.907048	1	XGboost	93739.438621	170856.270224
2	ADABOOST	381271.322241	409033.824130	2	ADABOOST	369010.449408	404865.582125
3	GradientBoosting	67651.546676	118738.580803	3	GradientBoosting	86008.052516	196024.777934
4	LGBM	49483.434268	131497.330684	4	LGBM	74560.860201	146477.497190
5	CatBoost	33753.696221	49243.290287	5	CatBoost	69720.873568	167151.332884



ELECCIÓN DEL MODELO

OVERFITTING?

PARÁMETROS GRIDSEARCH

- "**learning_rate**": ajustarla para evitar el sobreajuste.
- "**depth**": reducir profundidad del árbol
- "**L2_leaf_reg**": penalizar menos los valores atípicos
- Reducir "**subsample**" y "**colsample_bytree**" limitar la cantidad de datos que el modelo puede ajustar.



ELECCIÓN DEL MODELO



Optimizando los hiperparámetros,
El modelo no mejoró

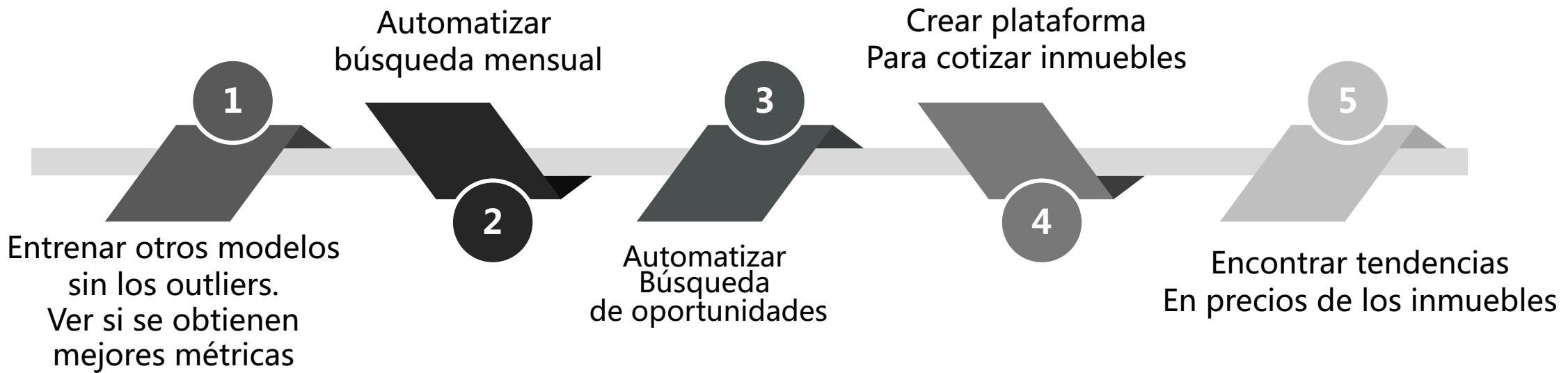
```
grid_search.best_params_  
  
{'colsample_bylevel': 0.9,  
 'depth': 5,  
 'l2_leaf_reg': 0.5,  
 'learning_rate': 0.05,  
 'subsample': 0.5}
```

```
RMSE TEST: 173745.00600245854  
MAE TEST: 71790.36032949104
```

4 CONCLUSIONES Y PRÓXIMOS PASOS



CONCLUSIONES Y PRÓXIMOS PASOS





GRACIAS

The BRIDGE