PREDICCIÓN DEL PRECIO DE COCHES DE SEGUNDA MANO

Mikel Guillen
mayo 2025

1. **Introducción**

Objetivo:

Predicción del precio de un coche de segunda mano mediante Machine Learning.

Público objetivo:

- Plataformas de compraventa
- Concesionarios
- Tasadores
- Particulares

Tipo de problema:

Regresión (el precio es una variable continua).

Evaluaremos en base a:

MAE (Mean Absolute Error)

2. **Recolección de datos y limpieza**

Dataset:

- Real obtenido de Datamarket con información de vendedores profesionales en España en 2023.
- 100.000 anuncios con 28 variables.

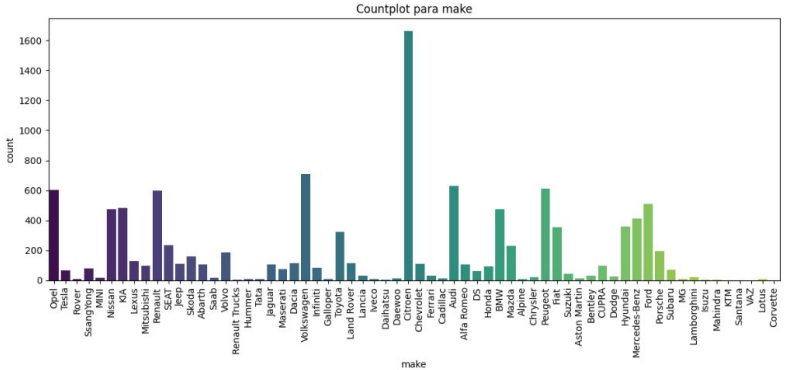
Limpieza de datos:

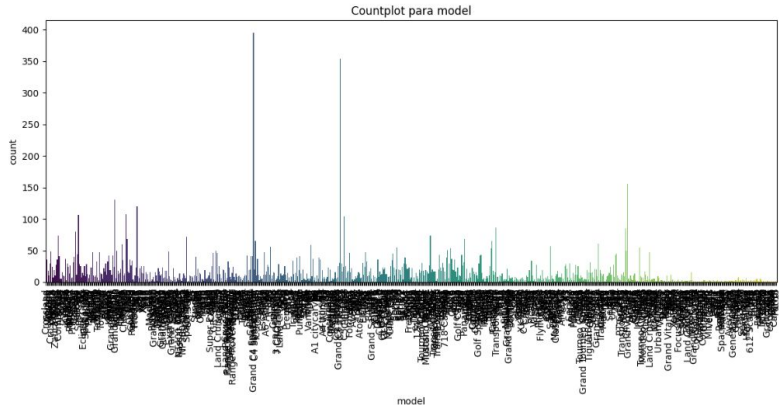
1. Eliminación de anuncios duplicados
2. Tratamiento de valores faltantes
3. Tratamiento de outliners con valores erróneos
4. Eliminación de variables que no aportan valor

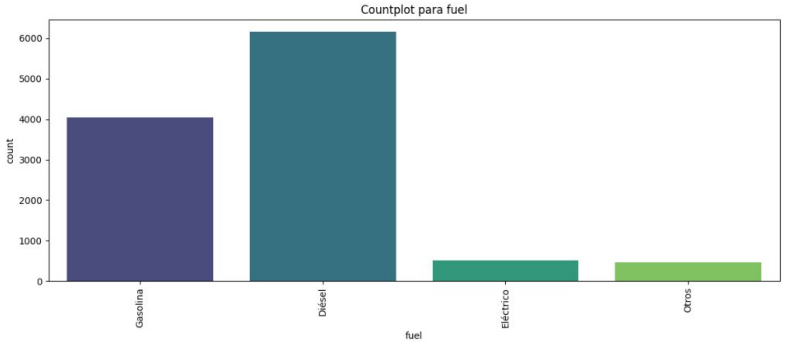
Resultado: 8 variables y 11185 anuncios

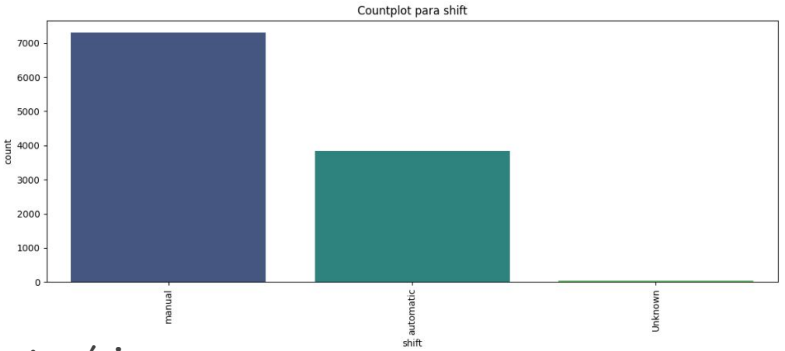
Tipo de Variable	Nombre de la Columna	Descripción
Numérica	year	Año de fabricación del vehículo
Numérica	kms	Kilometraje acumulado
Numérica	power	Potencia del motor (en CV o kW, según el caso)
Numérica	price	Precio del vehículo
Categórica	make	Marca del vehículo (ej. BMW, Audi, etc.)
Categórica	model	Modelo del vehículo (ej. A4, Golf, etc.)
Categórica	fuel	Tipo de combustible (ej. diesel, gasolina, eléctrico, híbrido)
Categórica	shift	Tipo de transmisión (manual, automática)

3. **Análisis exploratorio (EDA)**



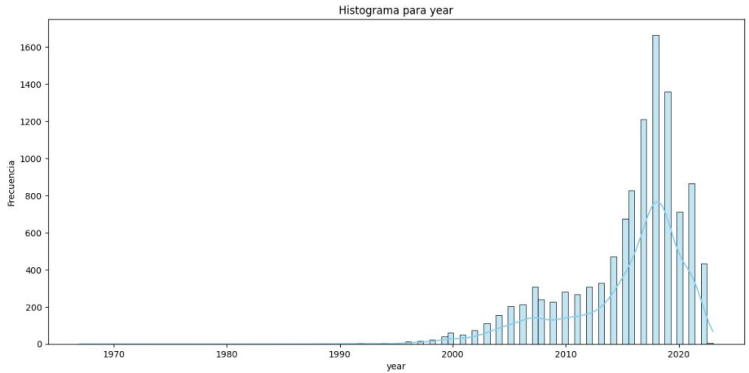


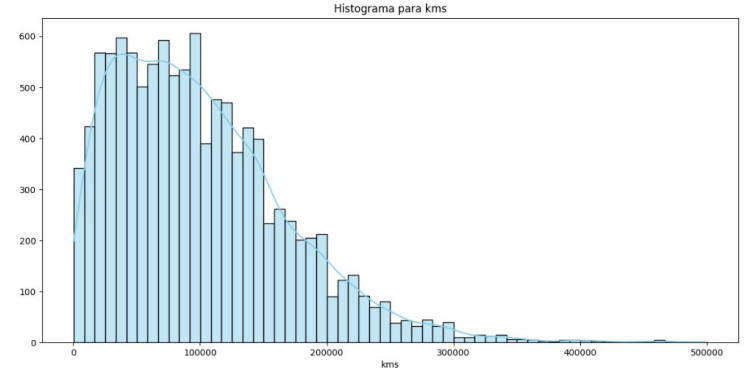


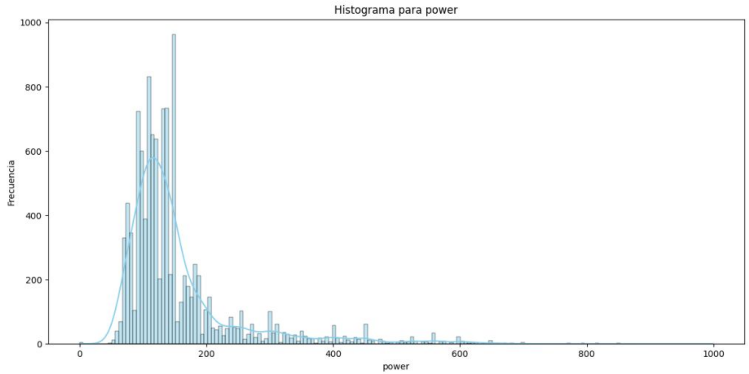


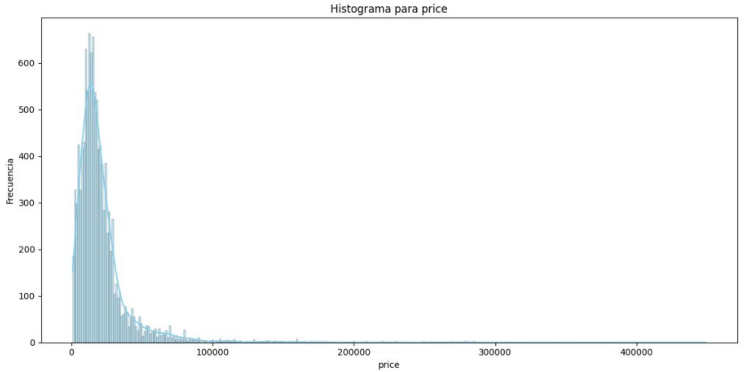
Variables categóricas

3. **Análisis exploratorio (EDA)**



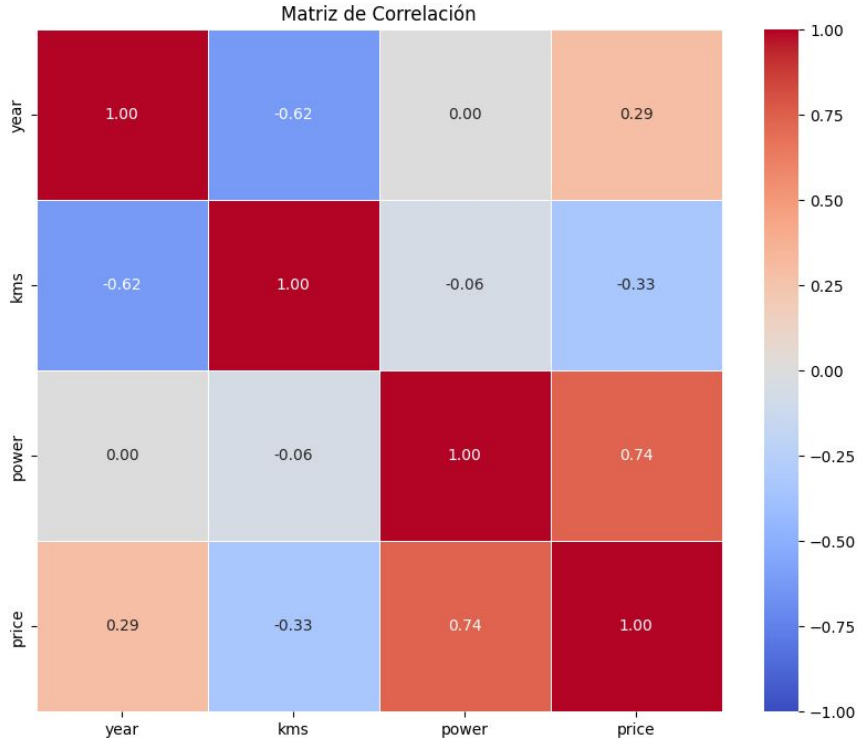


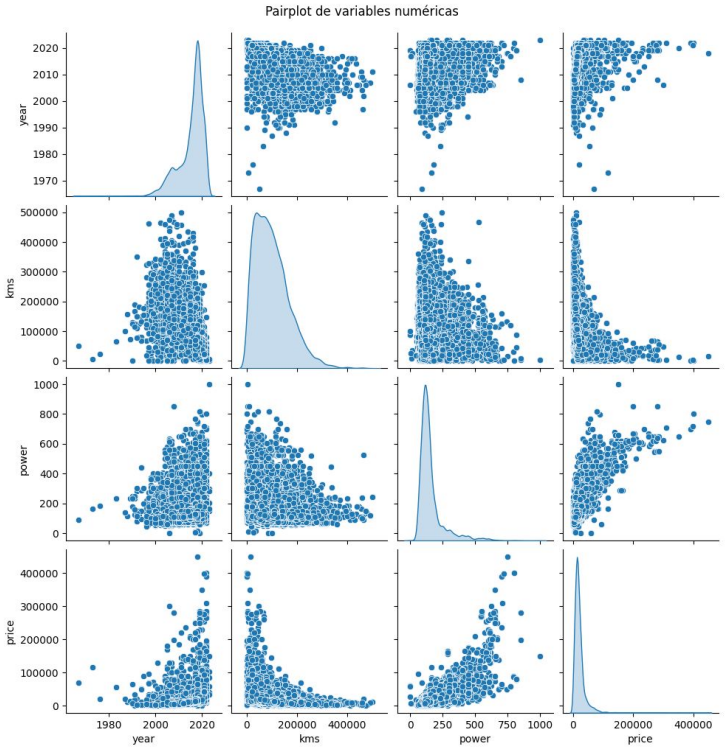




Variables numéricas

3. **Análisis exploratorio (EDA)**





4. **Preprocesamiento de datos**

Codificar variables categóricas:

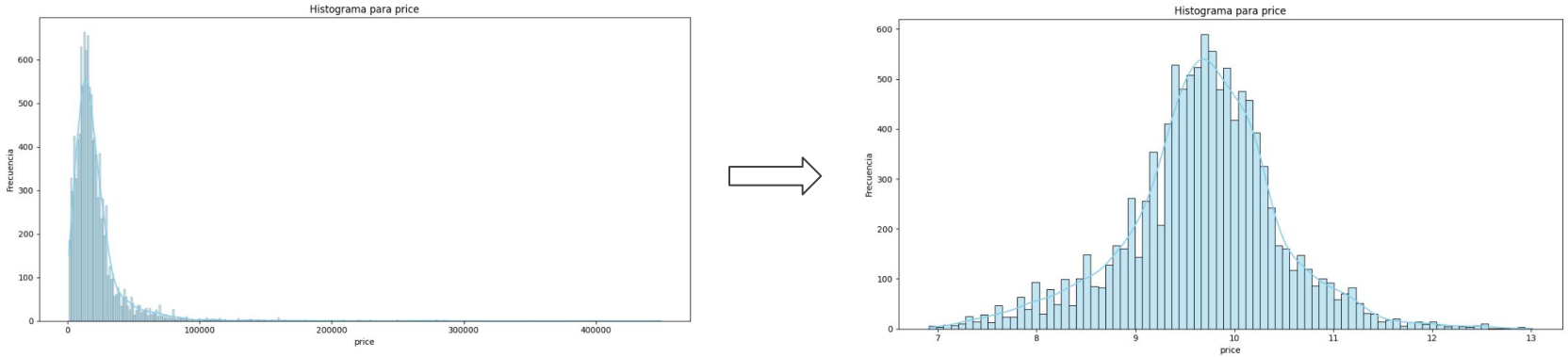
- Marca (64), tipo combustible (4) y cambio (3): one-hot encoding (71 variables).
- Modelos: 723 en total. Los clasificamos en 9 segmentos + otros y aplicamos one hot encoding.
    - Segmento A: Minicompactos
    - Segmento B: Coches pequeños
    - Segmento C: Compactos
    - Segmento D: coches grandes
    - Segmento E: coches de prestigio
    - Segmento F: coches de lujo
    - Segmento J: SUV
    - Segmento M: Familiares grandes
    - Segmento S: Deportivos
    - Otros

Total:

85 variables: 4 numéricas y 81 binarias

4. **Preprocesamiento de datos**

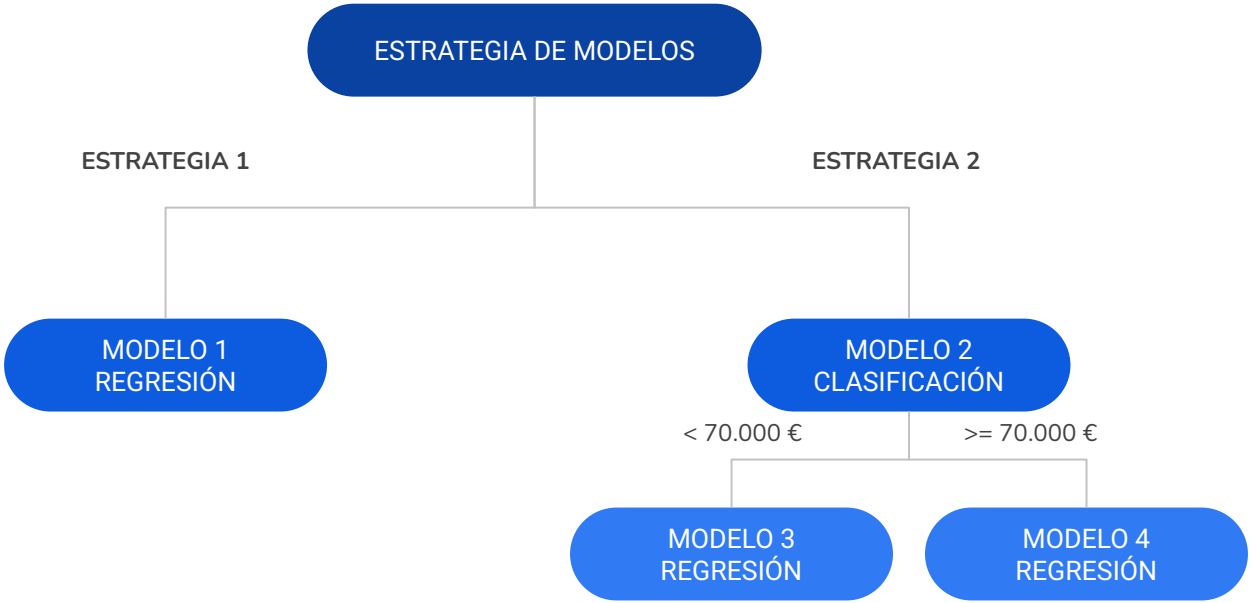
Transformación logarítmica del target precio.



Otras transformaciones desestimadas:

- Escalado: Los algoritmos utilizados no son sensibles al escalado
- PCA: No se han obtenido mejoras y se pierde interpretabilidad.

5. **Entrenamiento de modelos**

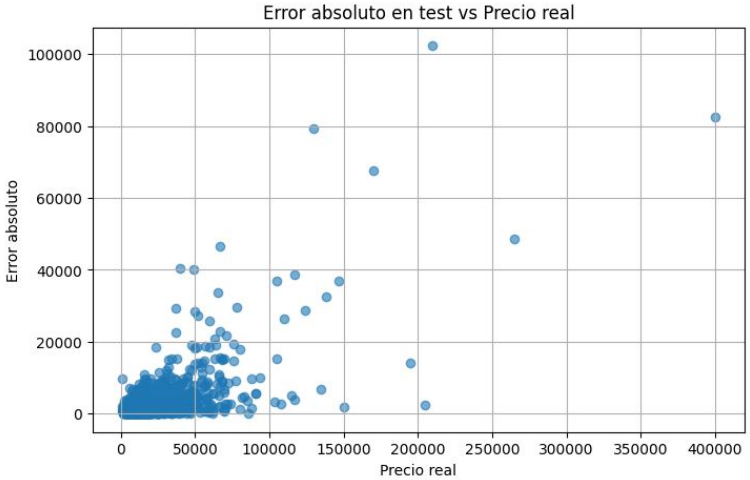


Nota: Se ha dividido el dataset 80% entrenamiento y 20% validación

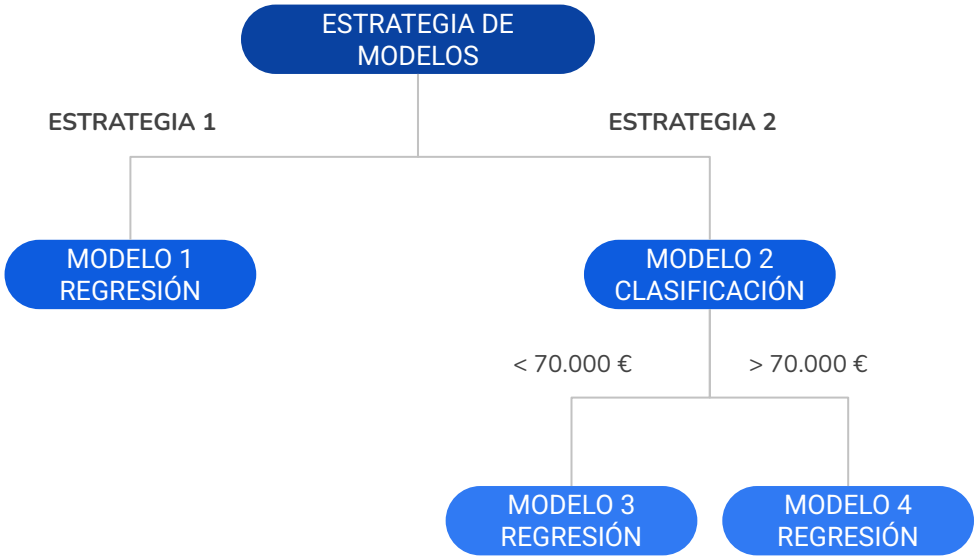
5. **Entrenamiento de modelos**

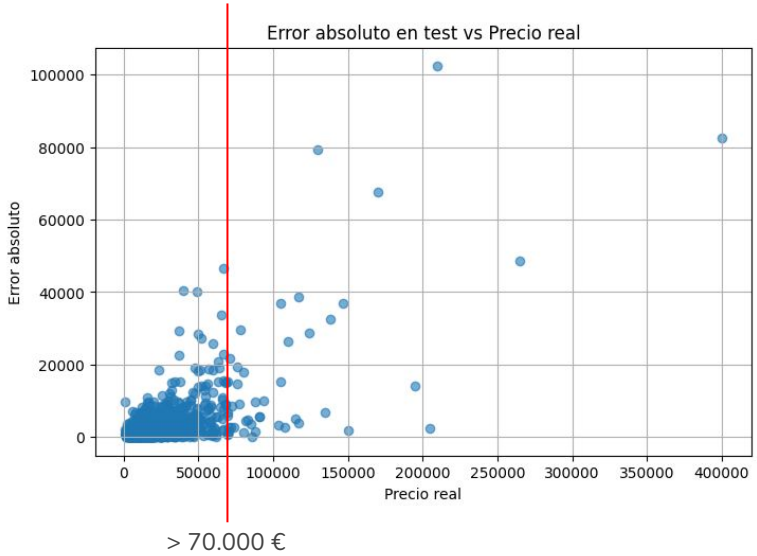
ESTRATEGIA 1 - MODELO 1: REGRESIÓN

#	Modelo	MAE Test	MAE Train	R^2 Test	R^2 Train
1	CatBoost	2746.74	2067.80	0.915	0.974
2	XGBoost	2861.58	1973.04	0.916	0.978
3	Random Forest	2903.76	1891.79	0.915	0.970
4	Gradient Boosting	2979.03	2085.75	0.889	0.978



5. **Entrenamiento de modelos**





5. **Entrenamiento de modelos**

ESTRATEGIA 2 - MODELO 2: CLASIFICACIÓN RANDOM FOREST

Matriz de Confusión

	Predicho: No premium (<70.000 €)	Predicho: Premium (≥70.000 €)
Real: No premium (<70.000 €)	1734	5
Real: Premium (≥70.000 €)	14	37

Reporte de Clasificación

Clase	Precisión	Recall	F1-score	Soporte
No premium (<70.000 €)	0.99	1.00	0.99	1739
Premium (≥70.000 €)	0.88	0.73	0.80	51

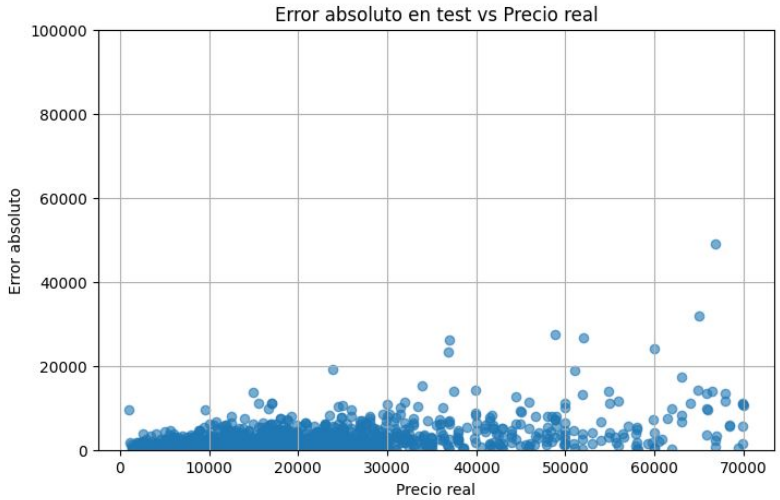
Métricas Generales

Métrica	Valor
Accuracy	0.99
Macro avg	Precisión: 0.94, Recall: 0.86, F1-score: 0.90
Weighted avg	Precisión: 0.99, Recall: 0.99, F1-score: 0.99

5. **Entrenamiento de modelos**

ESTRATEGIA 2 - MODELO 3: REGRESIÓN VEHÍCULOS NO PREMIUM

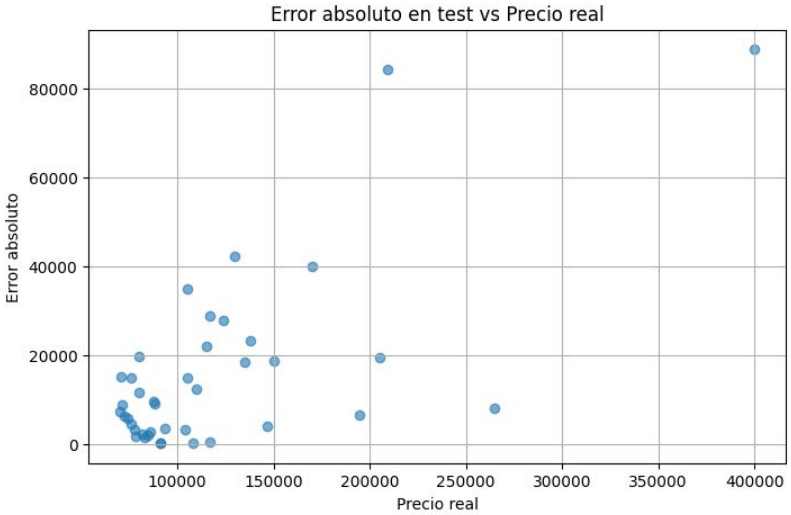
Nº	Modelo	MAE Test	MAE Train	R^2 Test	R^2 Train
1	CatBoost	2123.66	1245.45	0.921	0.977
2	XGBoost	2328.05	1636.32	0.912	0.960
3	Gradient Boosting	2362.26	1724.93	0.909	0.957
4	Random Forest	2464.73	1574.51	0.898	0.960



5. **Entrenamiento de modelos**

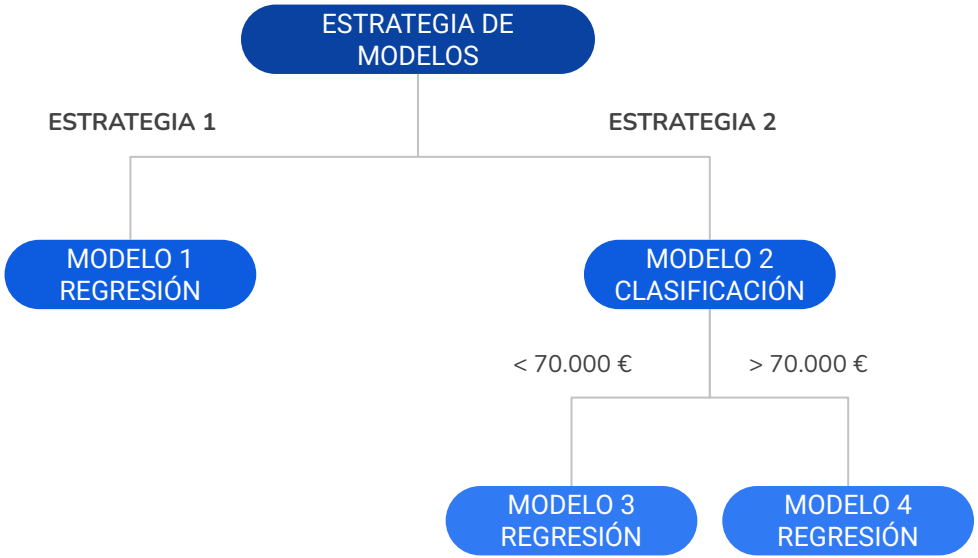
ESTRATEGIA 2 - MODELO 4: REGRESIÓN VEHÍCULOS PREMIUM

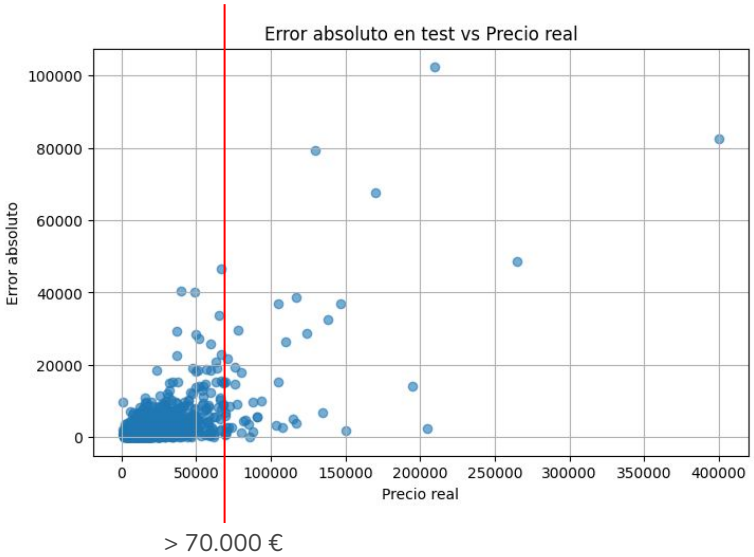
Nº	Modelo	MAE Test	MAE Train	R^2 Test	R^2 Train
1	Random Forest	15733.32	7854.14	0.840	0.958
2	Gradient Boosting	15825.78	6811.47	0.835	0.979
3	CatBoost	17785.09	6214.46	0.825	0.982
4	XGBoost	20509.03	15467.76	0.730	0.861



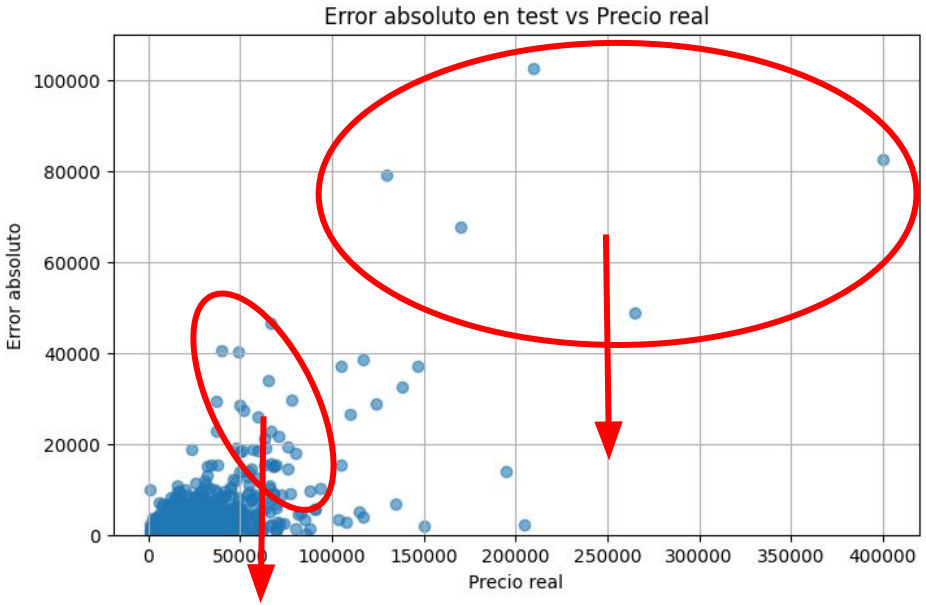
6. **Evaluación en Test**

RESULTADOS: ESTRATEGIA 1 VS ESTRATEGIA 2



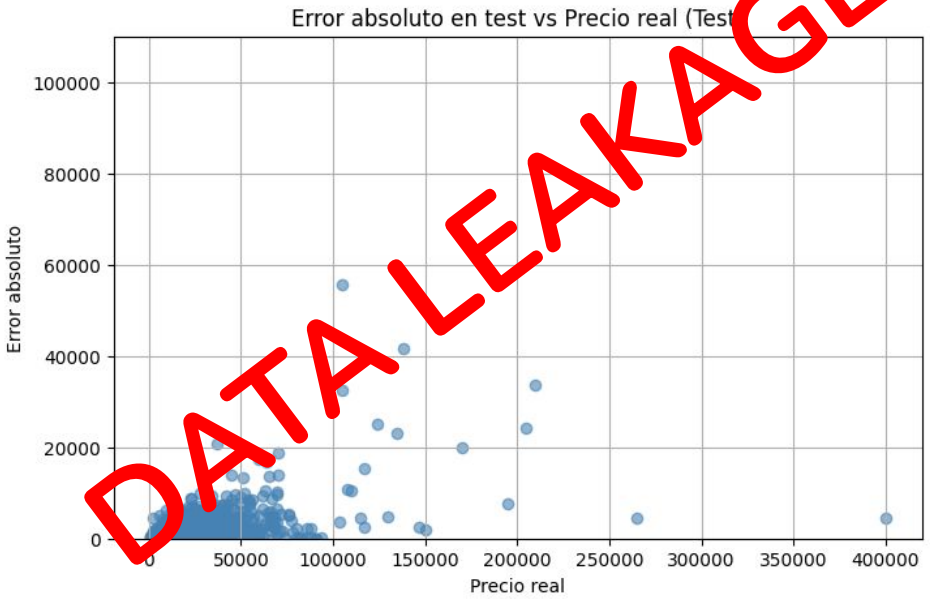


6. **Evaluación en Test**



ESTRATEGIA 1

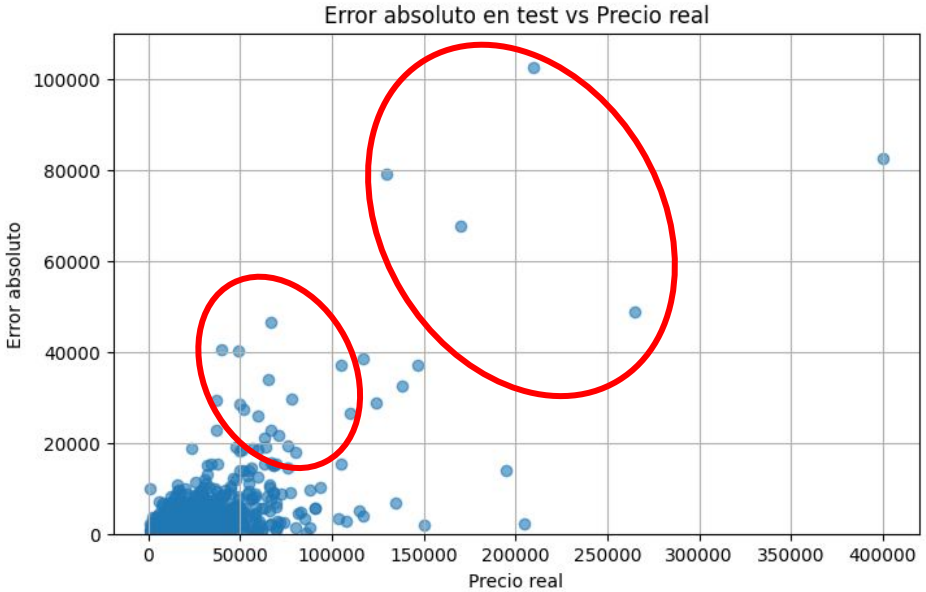
MAE TEST: 2746.74



ESTRATEGIA 2

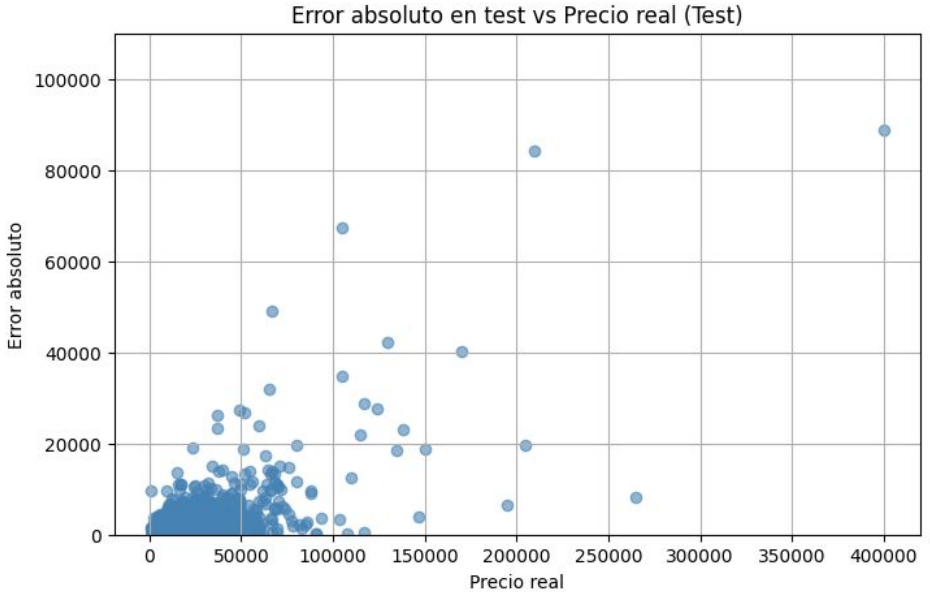
MAE TEST: 1586.83

6. **Evaluación en Test**



ESTRATEGIA 1

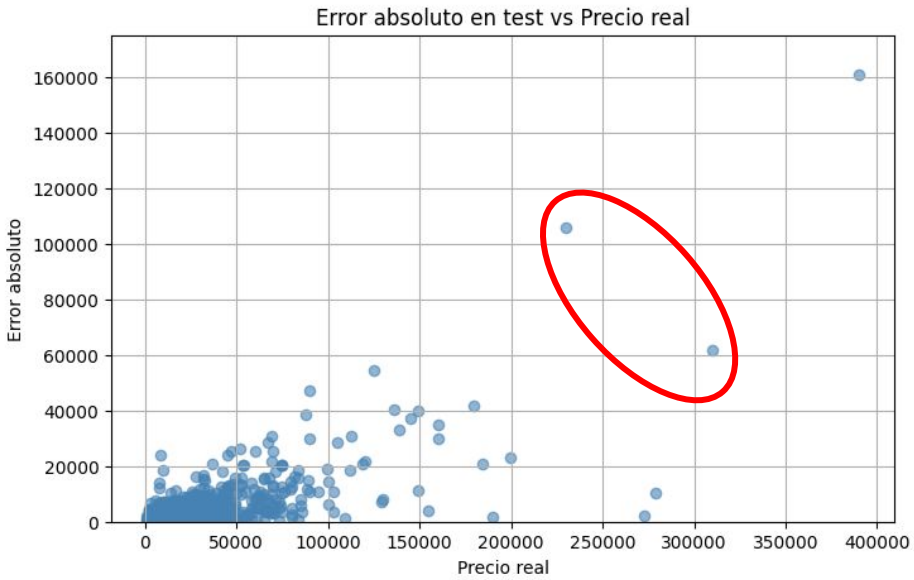
MAE TEST: 2746.74



ESTRATEGIA 2

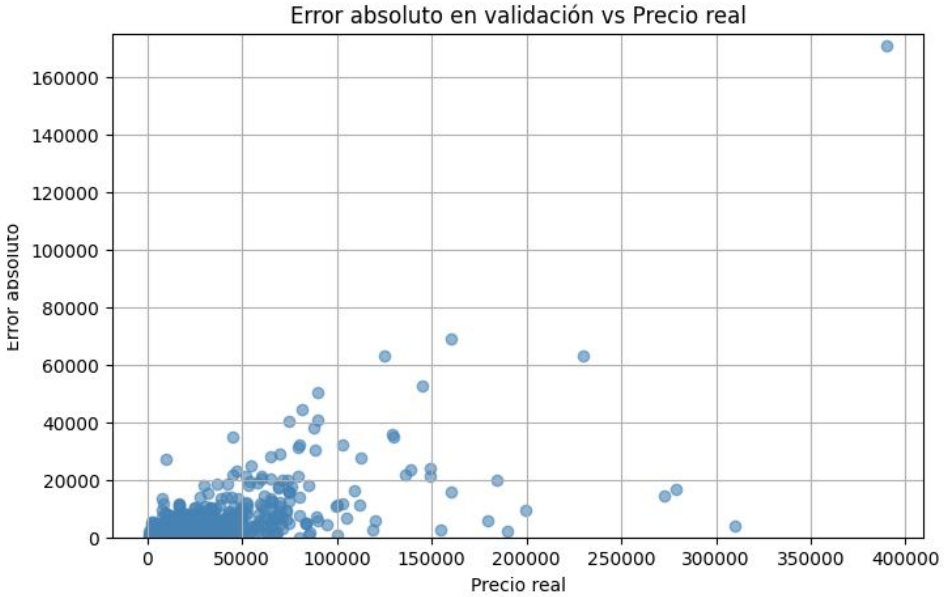
MAE TEST: 2457.47

7. **Validación**



ESTRATEGIA 1

MAE: 2752.98



ESTRATEGIA 2

MAE: 2.691,92

8. **Conclusiones**

- Los modelos dan una buena predicción del precio con un error aproximado del 12 % respecto al precio medio del vehículo.
- La estrategia 2 no obtiene mejoras sustanciales. Únicamente 256 vehículos premium en el dataset.

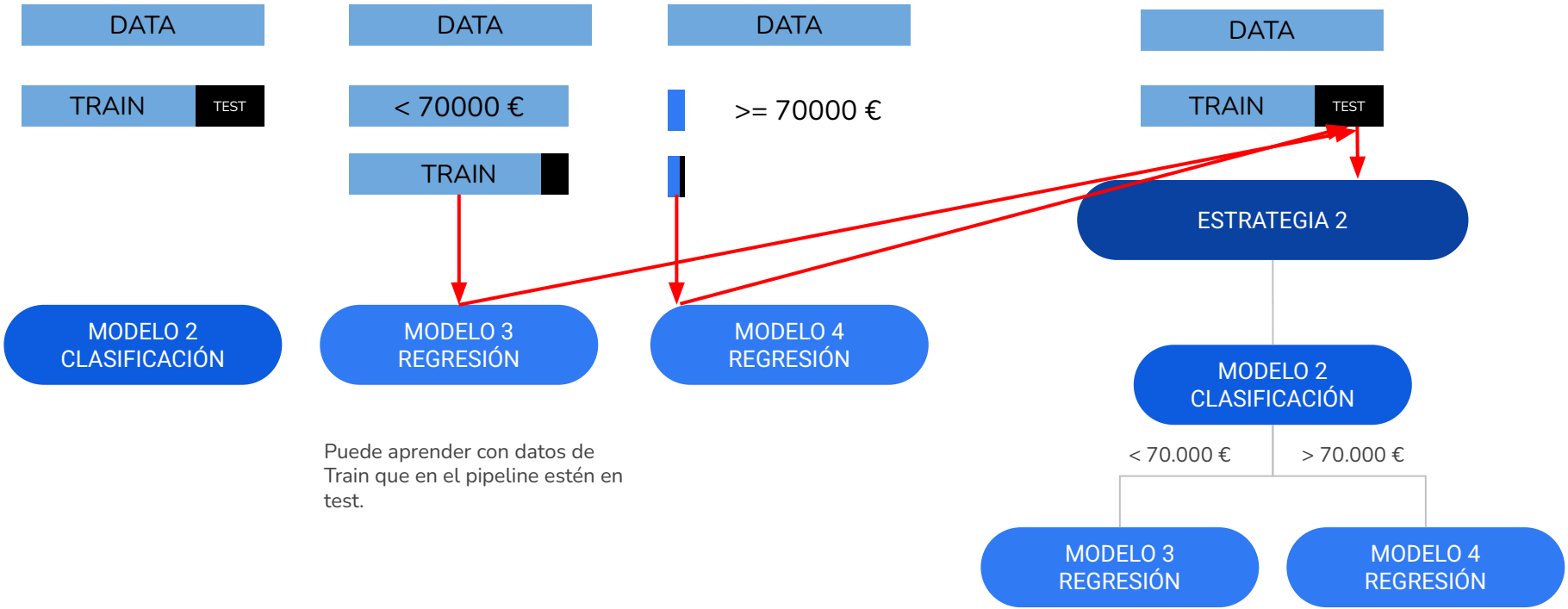
Estrategia	Conjunto	MAE (€)	% Error vs Media
Estrategia 1	Test	2.746,74	12,56 %
	Validación	2.752,98	13,01 %
Estrategia 2	Test	2.457,47	11,23 %
	Validación	2.691,92	12,72 %

9. **Futuros pasos**

- Entrenar los mejores modelos con todo el dataset
- Probar otro tipo de procesamiento de datos como target encoding:
  - Marca: substituir por el precio medio por marca
  - Segmento: substituir por el precio medio por marca y segmento
- Probar otros límites para la definición de la estrategia de modelos:
- Probar otro tipo de algoritmos como las redes neuronales.
- Conseguir más datos

ESKERRIK ASKO!!!

APÉNDICE



Puede aprender con datos de Train que en el pipeline estén en test.