Predicción de precio de casas a la venta

Proyecto Final - Machine Learning 2025

Nicolás, Diego y Bruno.

Equipo

Nicolás Fripp

Email:

nicolas.fripp@estudiantes.utec.com.uy

Diego Aguiar

Email:

diego.aguiar@estudiantes.utec.com.uy

Bruno Moraes

Email:

bruno.moraes@estudiantes.utec.com.uy

contexto

Base de datos:

- venta de casas en Uruguay.
- Mercado Libre y el Gallito, en los años 2023 y 2025.

Objetivo

Crear un modelo que logre **predecir el precio de las casas**.
Buscando cubrir las siguientes casuísticas de uso:

- Guia para marcar el precio de venta de una casa.
- Marcar precios sobre-valuadas o infra-valuadas. Marcando **oportunidades** de compra o de corrección de precios.
- Observar las variables más relevantes al momento de determinar el precio de una propiedad.

Dimensión del dataset

Observaciones

177.394

Variables

33

Variables

Float64 **#5**

Int64 **#3**

Object **#25**

Variables

- 0 NEIGHBORHOOD
- 1 COVERED_AREA
- 2 LISTING_TYPE_ID
- 3 ADDRESS_CITY_NAME
- 4 ROOMS
- 5 WITH_VIRTUAL_TOUR
- 6 HAS_AIR_CONDITIONIN
- 7 PROCESS_DATE
- 8 TOTAL_AREA
- 9 DESCRIPTION
- 10 ITEM_CONDITION
- 11 PROCESS_DATE.1
- 12 ADDRESS_STATE
- 13 ADDRESS
- 14 CONDITION
- 15 TITLE
- 16 PRICE
- 17 ADDRESS_LINE
- 18 CATEGORY_ID
- 19 ORIGEN
- 20 SITE_ID
- 21 HAS_TELEPHONE_LINE
- 22 ID
- 23 DESCRIPTION.1
- 24 FULL_BATHROOMS
- 25 OPERATION
- 26 BEDROOMS
- 27 CURRENCY_ID
- 28 PROPERTY_TYPE
- 29 POSITION
- 30 ARTICLE_ID
- 31 GARAGE
- 32 CONSTRUCTION_YEAR

Origen de datos

MELI

mercado libre

Gallito

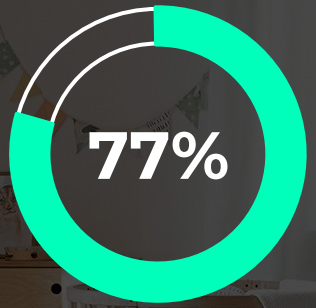
gallitoluis


Limpieza
de datos

Eliminamos las siguientes variables:

-'DESCRIPTION'
- 'TITLE'
- 'CATEGORY_ID'
- 'SITE_ID', 'ID'
-'DESCRIPTION.1'
-'OPERATION',
-'PROPERTY_TYPE'
-'ADDRESS' -'ADDRESS_LINE'
-'ROOMS'
-ARTICLE_ID

Datos Duplicados



Tenemos 136.560 datos duplicados

Propiedades repetidas por varias cargas de datos, e iguales precios.

Merge de datos

Variables de MELI	Unificamos datos	Variables del Gallito
NEIGHBORHOOD		ADDRESS_CITY_NAME
PROCESS_DATE		PROCESS_DATE.1
CONDITION		ITEM_CONDITION

Chequeos de datos Nulos y tratamiento

- Reagrupamos “condicion”: usado, nuevo , sin especificar, en construccion.
- Reagrupamos los departamentos en: Montevideo, Maldonado, Canelones y Otros.
- Adecuamos los precios en base a la fecha de imputación y el tipo de cambio del momento.

Imputación de datos: Departamento

- Dado el barrio o ciudad, imputamos el departamento.
  - Moda
  - Las restantes manualmente

```
"montoya": "Maldonado",
"la residence": "Maldonado",
"tio tom": "Maldonado",
"fray marcos": "Florida",
```

	NEIGHBORHOOD	ADDRESS_STATE
37	Carrasco	Montevideo
99	Carrasco	Montevideo
102	Carrasco	Montevideo
158	Carrasco	Montevideo
235	Carrasco	Montevideo
...	...	...
177283	Carrasco	NaN
177288	Carrasco	NaN
177300	Carrasco	NaN
177347	Carrasco	NaN
177353	Carrasco	NaN

Imputación de datos: Covered_area

Imputamos **covered_area =1** en base al área promedio por grupos de cantidad de cuartos y baños.

Imputación de datos: listing_type_id

Dato que venía de MELI.
Por ende todos los de gallito son nulos:

- imputamos estos con “gallito”

Análisis de la variable Moneda

Todo USD

Las observaciones en UYU, las analizamos y corresponden a USD

	PRICE	CURRENCY_ID
121108	16500.00	$U
38920	17000.00	$U
115007	17000.00	$U
98394	17000.00	$U
78158	17000.00	$U
119440	17000.00	$U
62513	17000.00	$U
112050	18000.00	$U
109760	18000.00	$U

Análisis de la variable objetivo: precio

- Eliminación de observaciones con patrón “111”
- Eliminamos 5 observaciones con precios menores a 1000 USD

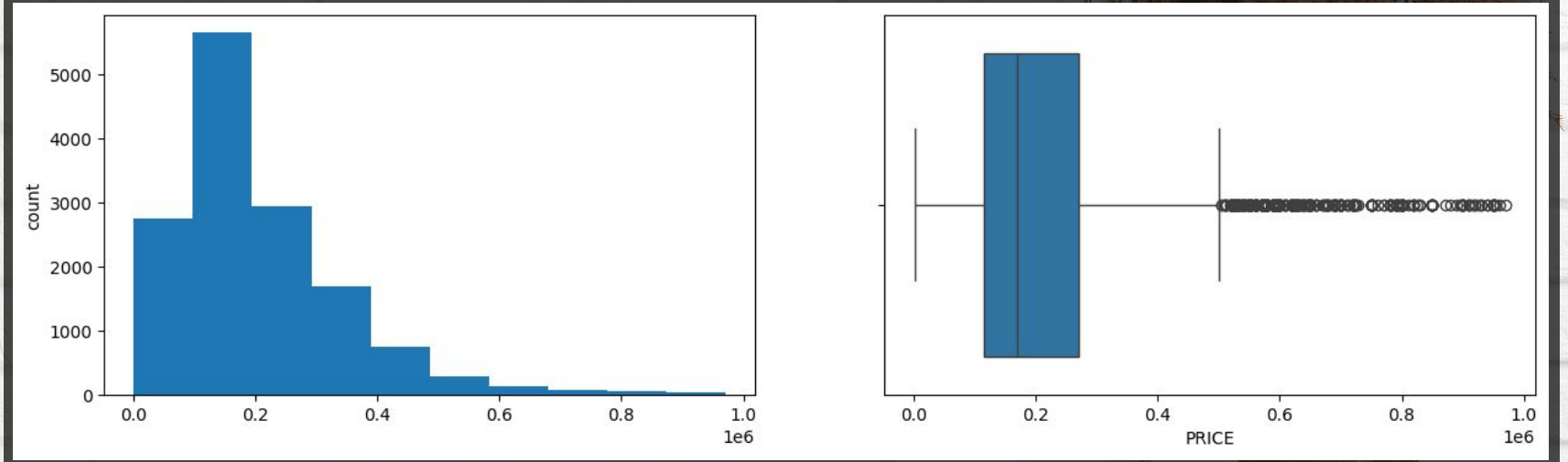
Análisis exploratorio

Univariado y Bivariado

Muestra Final

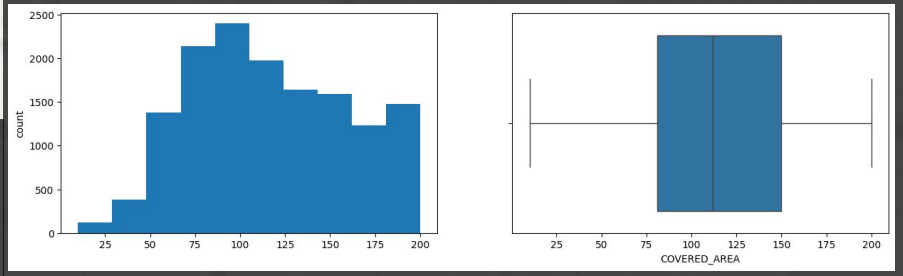
- Cuenta con 14.337 observaciones.
- Una variable objetivo **precio**.
- Once variables explicativas: Barrio, Área cubierta (m2), tipo de publicación, si tiene tour virtual, si tiene aire acondicionado, departamento, condición, origen, baños, dormitorios, garages.

Univariado - Variable Objetivo

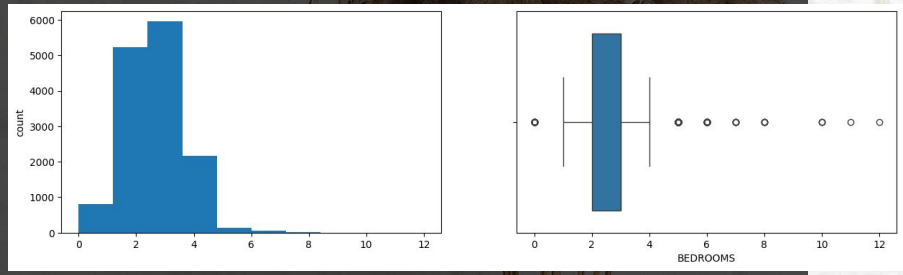


Univariado - Explicativas Numéricas

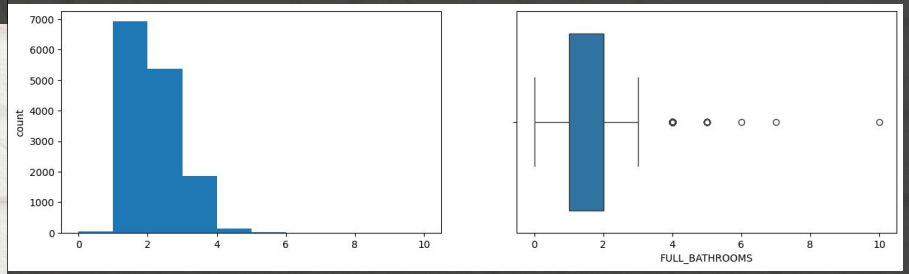
covered_area



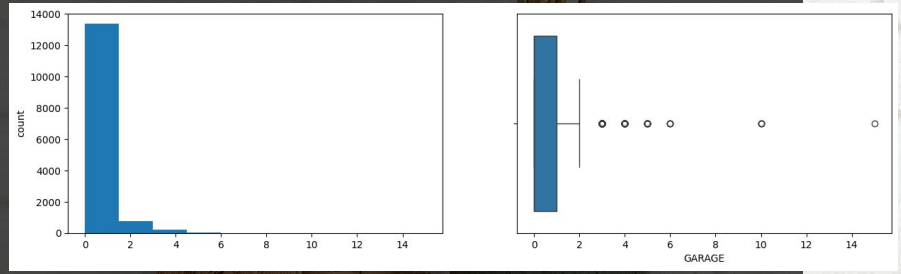
bedrooms



bath

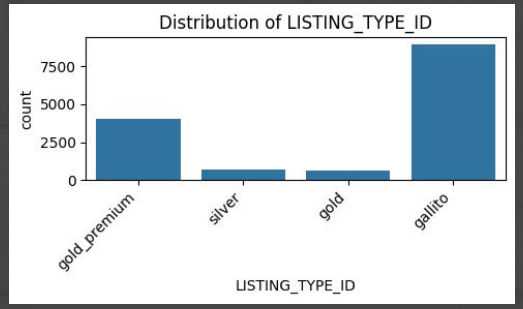


garage

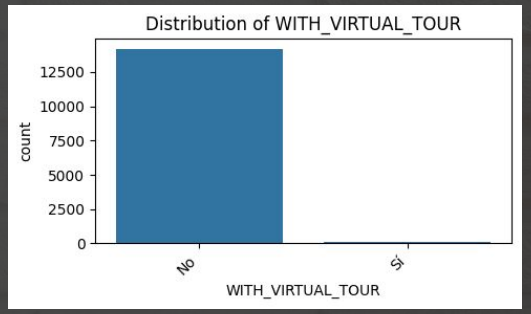


Univariado - Explicativas Categóricas

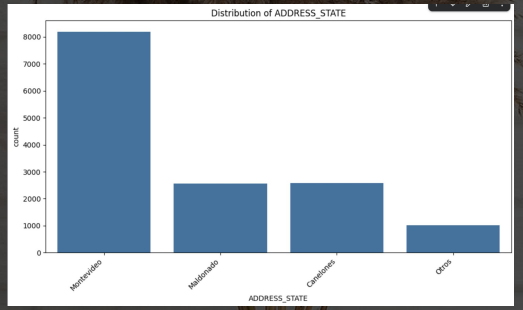
listing_type



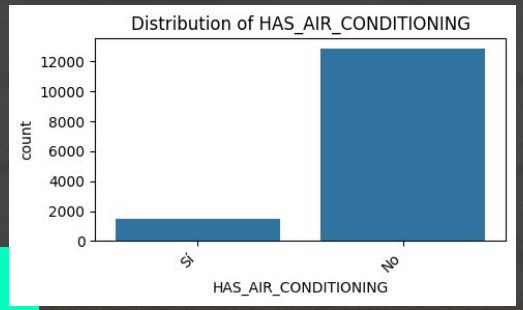
virtual_tour



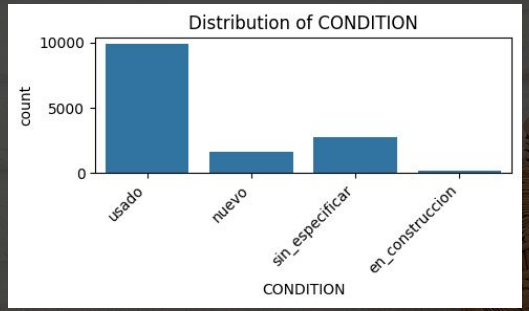
address_state



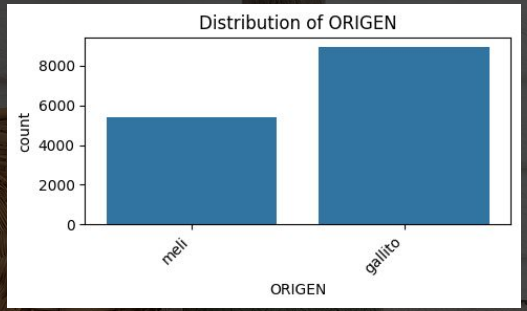
air_conditioning



condition

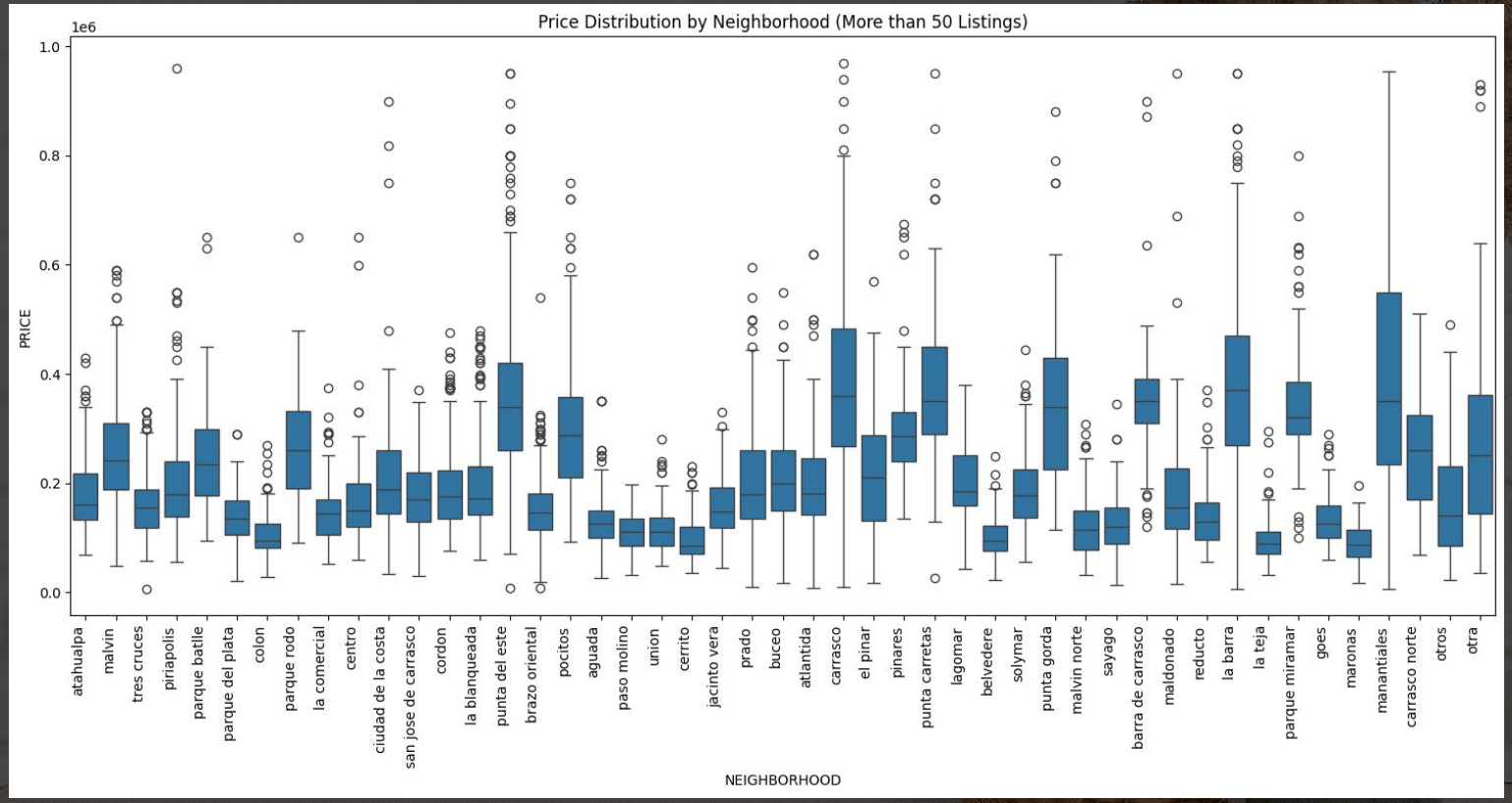


origen



Bivariado - Mapa de Calor

Bivariado - Diagrama de Caja




MODELOS

Regresión Lineal - Introducción

- Elegimos regresión lineal como modelo base: simple, rápida y fácil de interpretar.
- Variable objetivo: log(precio).
- Variables explicativas:
  - m^2 también en log.
  - Variables categóricas → dummies.
  - Se usan las variables transformadas del análisis exploratorio.

Regresión Lineal - Medida de Éxito

- Métrica principal: MAPE.
- Indica el error porcentual promedio.
- MAPE = 50% → predicciones se desvían ~50% del valor real.

Regresión Lineal - Resultados

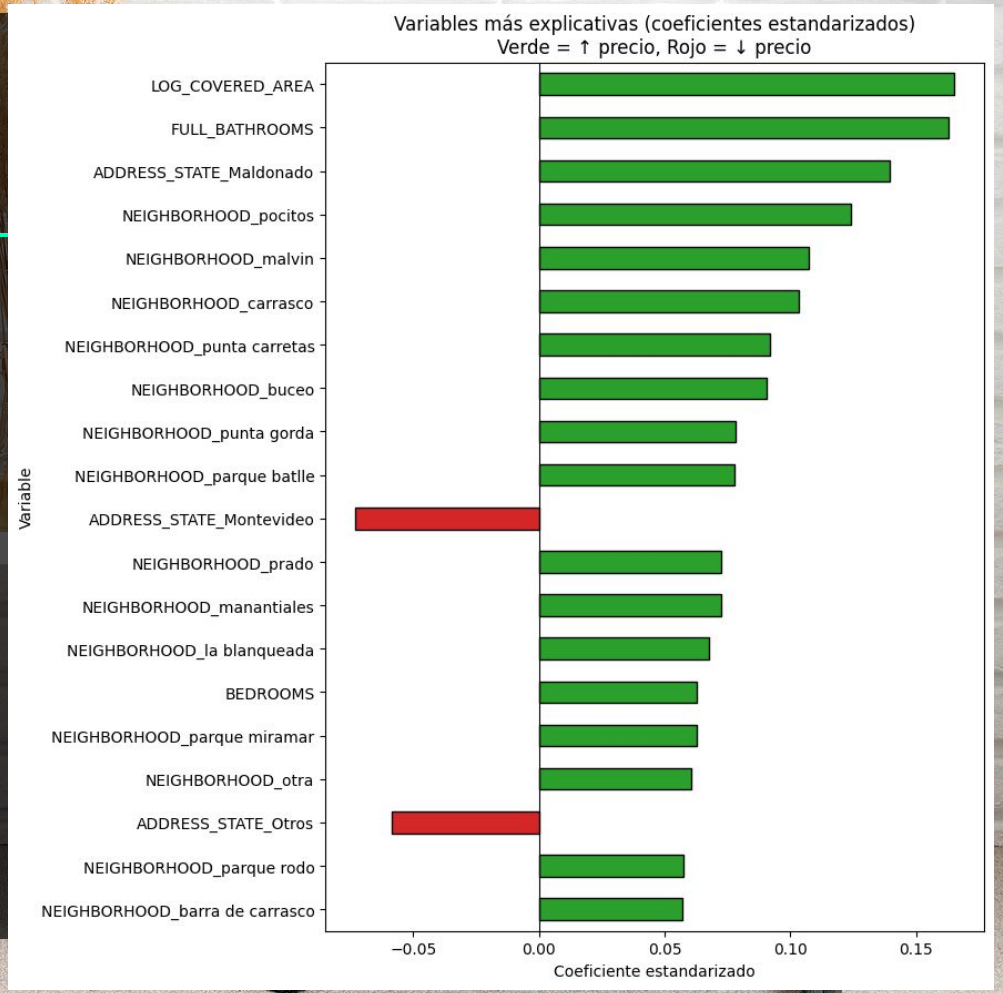
- R^2 = 0.52 → explica ~52% de la variación del precio.
- MAPE = 40% → error típico: ± USD 80.000 (promedio = 200K).
- Buen baseline, pero limitada capacidad predictiva.

Se necesitan modelos más flexibles.

```
[EVALUACIÓN OLS (TEST - Escala Original)]
  R²   = 0.5216
  RMSE = 92,933.08
  MAE  = 57,197.59
  MAPE = 39.90%
```

Regresión Lineal - Explicativas

- Mayor impacto: m^2, baños, barrios específicos.
- Efectos negativos:Montevideo y departamentos que no son ni Maldonado ni Canelones.
- Se usan coeficientes estandarizados para comparar importancia.



Regresión Lineal - Regularización

- Probamos métodos de regularización:
  - Ridge (L2): reduce la magnitud de los coeficientes.
  - Lasso (L1): penaliza coeficientes y puede dejar algunos en cero (selección de variables).

- Objetivo: reducir sobreajuste y mejorar predicción.

Regresión Lineal - Resultados de la Regularización

- R^2 y MAPE similares al modelo lineal.
- Esperable porque:
  - pocas variables,
  - baja multicolinealidad,
  - todas aportan información.
- La regularización no agrega valor en esta base.

```
--- REGRESIÓN RIDGE (L2) ---
Alpha óptimo Ridge: 0.5857
[EVALUACIÓN RIDGE (TEST - Escala Original)]
  R²    = 0.5212
  RMSE = 92,974.29
  MAE   = 57,208.72
  MAPE = 39.93%
```

```
--- REGRESIÓN LASSO (L1) ---
Alpha óptimo Lasso: 0.0003
[EVALUACIÓN LASSO (TEST - Escala Original)]
  R²    = 0.5149
  RMSE = 93,584.11
  MAE   = 57,403.03
  MAPE = 40.22%
```

Modelo: Gradient Boosting

CatBoost

De las 10 variables explicativas 6 son categóricas

Categóricas

NEIGHBORHOOD, LISTING_TYPE_ID, ADDRESS_STATE, CONDITION, WITH_VIRTUAL_TOUR & HAS_AIR_CONDITIONING

Eliminamos:

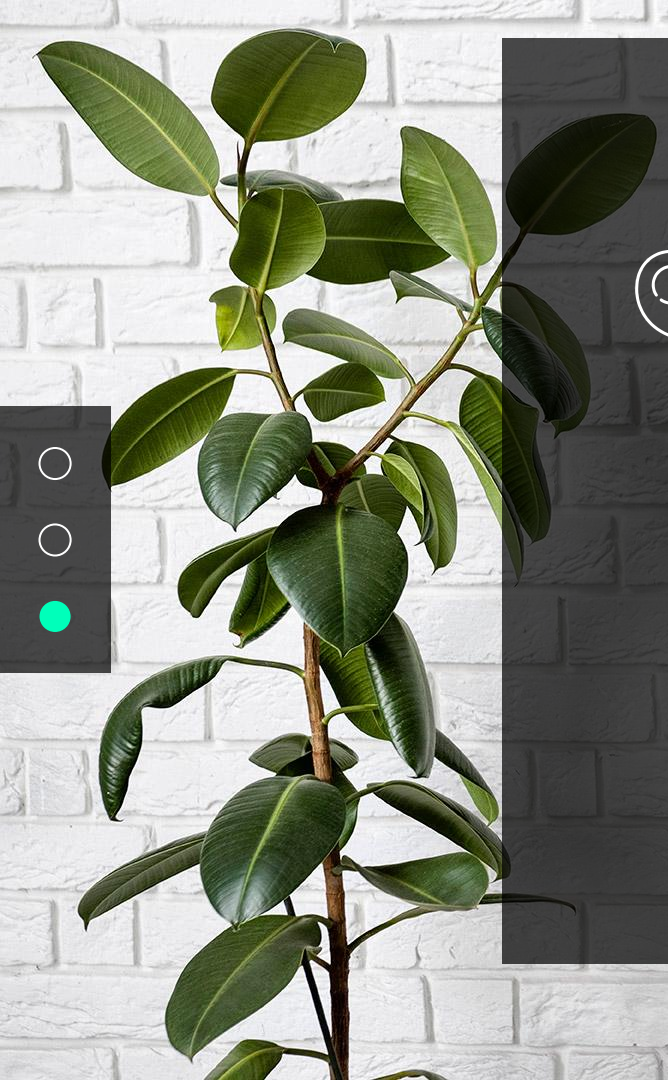
La variable Origen (data similar a LISTING_TYPE_ID)

Bases sin outliers

Dataset con barrios menores unificados

Métrica

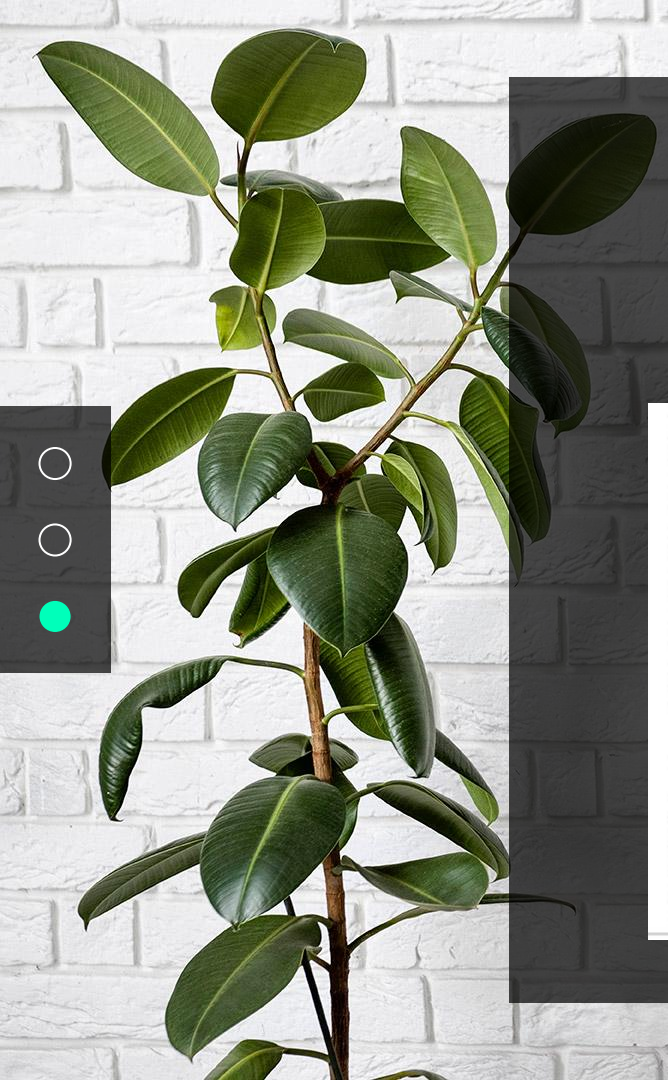
MAPE



Modelo: CatBoost

NEIGHBORHOOD & ADDRESS_STATE

Probamos sin las agrupaciones, y nos dio mejores resultados, por lo tanto los dejamos como venían de origen en el dataset.



Modelo: CatBoost

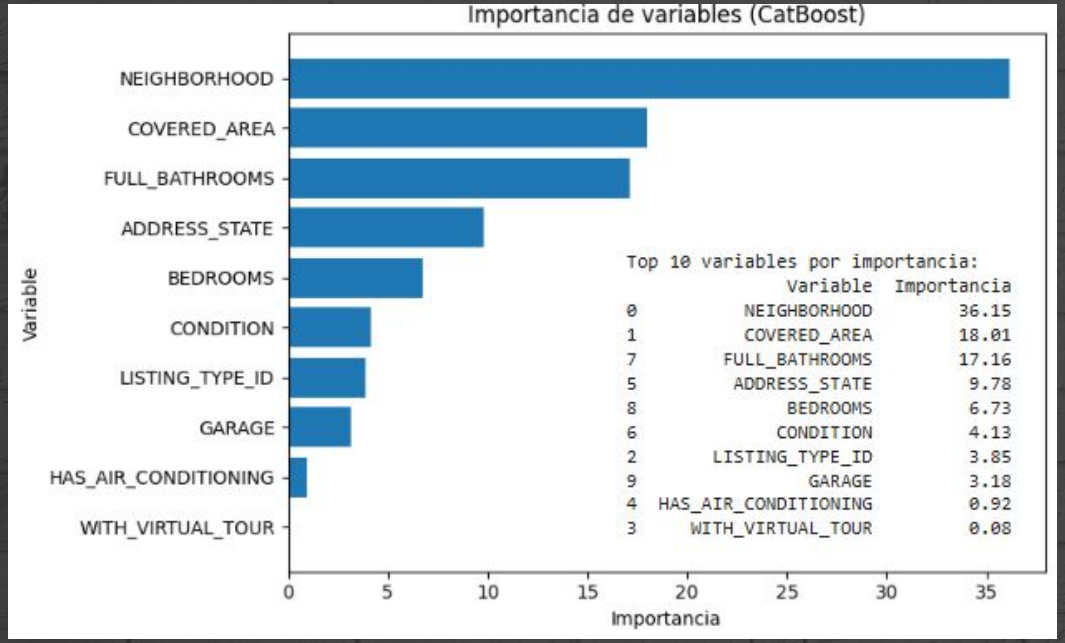
Resultados:

```
Starting CatBoost Cross-Validation...
Fold 1 - RMSE: 78,627.92, MAE: 49,047.67, R²: 0.653, MAPE: 34.59%
Fold 2 - RMSE: 73,774.59, MAE: 49,183.64, R²: 0.702, MAPE: 40.99%
Fold 3 - RMSE: 72,372.11, MAE: 48,771.51, R²: 0.697, MAPE: 30.88%
Fold 4 - RMSE: 67,871.67, MAE: 46,583.43, R²: 0.710, MAPE: 27.94%
Fold 5 - RMSE: 74,020.09, MAE: 49,587.11, R²: 0.684, MAPE: 31.66%
Fold 6 - RMSE: 73,365.28, MAE: 49,297.60, R²: 0.689, MAPE: 30.47%
Fold 7 - RMSE: 71,464.20, MAE: 48,136.16, R²: 0.684, MAPE: 33.89%
Fold 8 - RMSE: 88,084.09, MAE: 54,432.44, R²: 0.634, MAPE: 40.11%

--- Cross-Validation Results ---
Average RMSE: 74,947.49 ± 5,697.46
Average MAE: 49,379.94 ± 2,105.47
Average R²: 0.682 ± 0.024
Average MAPE: 33.82%
```

Modelo: CatBoost

Importancia de las variables:

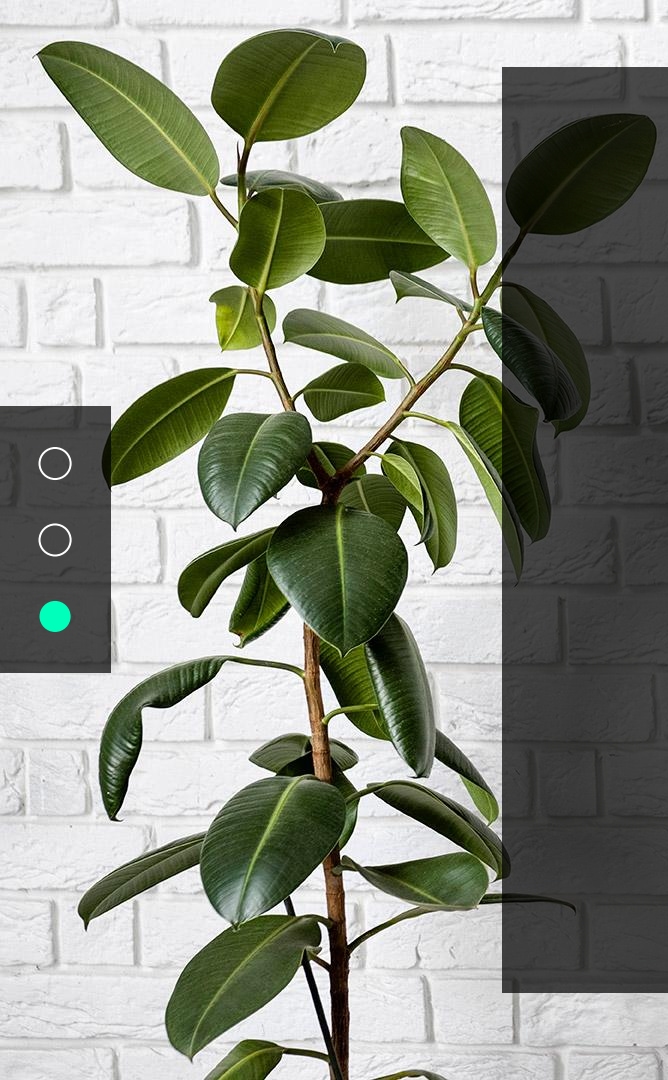


Catboost + Optuna

Utilizamos Optuna para optimizar los hiperparámetros de CatBoost.

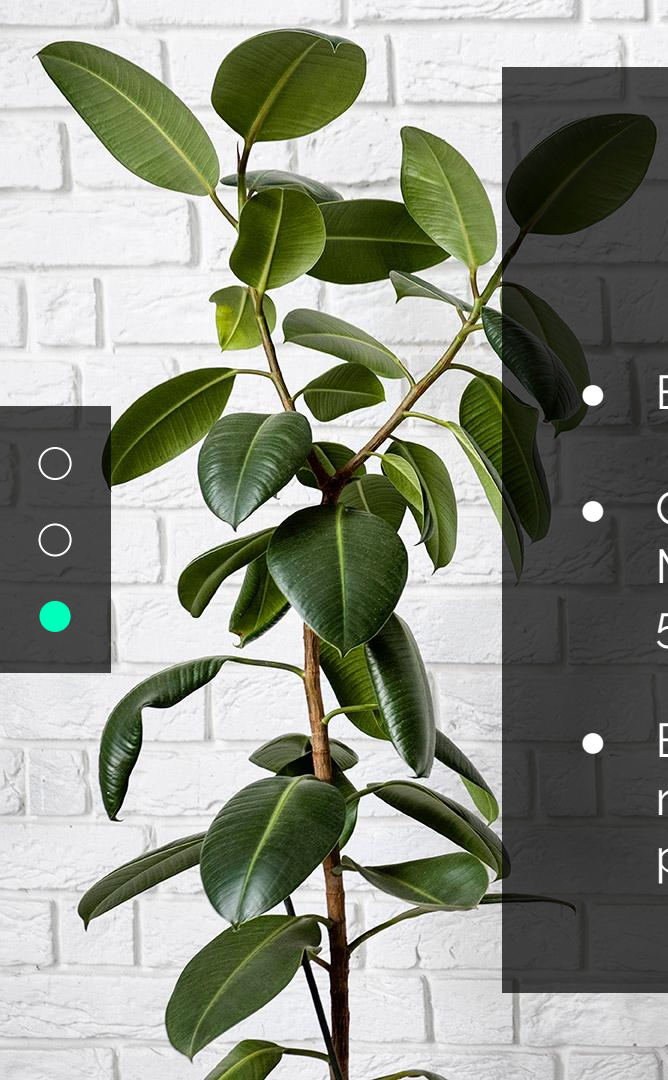
Resultado:

```
=== Mejor Resultado de Optuna ===
Mejor MAPE Promedio (K-Fold): 32.01%
```



Resultados

	MAPE
RL	39.9 %
Catboost	33.8%
Catboost + optuna	32%
Catboost + optuna para Montevideo	25%
Catboost + optuna para Canelones	26%
Catboost + optuna para Maldonado	37%



Conclusión

- El mejor modelo es CatBoost optimizado
- Con un precio promedio de USD 200.000, un MAPE de 25% implica un error típico de ± USD 50.000 por vivienda en Montevideo.
- Esto lo convierte en una herramienta útil como referencia inicial de precios, no como valuación profesional


Gracias