

# PRUEBA TÉCNICA MERCADO LIBRE

*Data Scientist: Albert Alejandro Durán Toledo*

*17 de Julio del 2025*

## Introducción

El objetivo de este proyecto es desarrollar un modelo de machine learning que permita predecir cuántas unidades puede vender un artículo nuevo publicado en una plataforma de e-commerce, utilizando para ello datos históricos de publicaciones anteriores.

En una primera etapa, se planteó como objetivo predecir las ventas por item\_id, sin embargo, tras realizar un análisis exploratorio del dataset, se identificó que cada fila representa una única publicación de un artículo, es decir, no se cuenta con un historial temporal de ventas por producto. Esto llevó a replantear el enfoque del proyecto para alinear los objetivos con la estructura real de los datos.

## Objetivo final

Predecir la variable sold\_quantity, entendida como el número de unidades vendidas por publicación de un artículo nuevo. El objetivo es estimar, al momento de publicar un artículo nuevo, cuántas unidades se espera que venda, considerando atributos como el precio, la modalidad de envío, características del vendedor, descuentos aplicados, fecha de publicación y otras variables disponibles.

Este tipo de predicción es altamente relevante para áreas como Marketing, Comercial y Business Intelligence, ya que puede utilizarse para anticipar el desempeño de nuevas publicaciones, mejorar la segmentación de productos, personalizar recomendaciones para vendedores, y diseñar estrategias de precios o envíos más efectivas que impulsen la conversión y aumenten el volumen de ventas.

### 1. Análisis Exploratorio de Datos (EDA)

Para comenzar el análisis, se desarrolló una clase DataAnalyzer que nos permitió explorar y preparar el dataset de manera estructurada. El objetivo inicial fue entender la calidad de los datos y evaluar su utilidad para una tarea de predicción.


#### Diagnóstico de calidad

- Se detectaron valores faltantes principalmente en variables como price, base\_price, attributes, pictures y warranty.

- Se eliminaron columnas irrelevantes o no útiles para modelado como tags, variations, pictures, seller\_country, seller\_id, entre otras, por contener información poco estructurada o identificadores sin valor predictivo.
- Se encontraron y corrigieron tipos de datos incorrectos, como fechas mal parseadas y flags (shipping\_is\_free, etc.) que no eran booleanos.
- Se verificaron duplicados, eliminando aquellos registros idénticos.
- Se detectaron inconsistencias como precios negativos, cantidades negativas y fechas de publicación excesivamente antiguas.
- Se identificaron outliers en el precio, con métodos como Z-Score e IQR. Finalmente, para evitar eliminar grandes volúmenes de datos, se optó por reemplazar los precios outliers por el percentil 95, preservando la distribución general.

## 2. Feature Engineering:

Se implementó la clase FeatureEngineer, la cual generó nuevas variables y seleccionó las más útiles para el modelo de predicción.

-  Variables creadas:
- discount\_pct: proporción de descuento entre base\_price y price.
- is\_discounted: flag que indica si hubo descuento.
- days\_since\_created: días desde la publicación hasta el análisis.
- Además, se convirtieron variables booleanas a enteros (0 y 1) para compatibilidad con modelos como LightGBM y XGBoost.

## 3. Modelado Predictivo

-  Tarea de ML definida:

Predecir la variable sold\_quantity, entendida como el número de unidades vendidas por artículo nuevo. Se busca estimar cuántas unidades venderá una publicación de un artículo nuevo, con base en sus atributos, fecha, precio, y tipo de envío.

-  División del dataset:

Se dividió el dataset en conjunto de entrenamiento (70%) y prueba (30%), con estricto control del random\_state para reproducibilidad.

-  Modelos entrenados:

CatBoost Regressor

LightGBM Regressor


XGBoost Regressor

Se usó GridSearchCV para optimizar hiperparámetros con validación cruzada de 3 folds. Los modelos fueron evaluados con las métricas: RMSE, MAE y MAPE.


## 4. Análisis de Resultados

-  Métricas en entrenamiento (valores promedio):

Modelo	RMSE	MAE	MAPE (%)
XGBoost	2.16	1.19	70.97
LightGBM	2.23	1.23	72.51
CatBoost	2.55	1.47	74.54

 Métricas en el conjunto de prueba:




Modelo	RMSE	MAE	MAPE (%)
XGBoost	2.47	1.35	79.04
LightGBM	2.44	1.34	76.88
CatBoost	2.59	1.49	74.41

-  Elección del modelo:
  - Elegimos LightGBM como el modelo ganador por tener el mejor RMSE y MAE en dataset de testeo.
  - El gap entre métricas de entrenamiento y prueba es pequeño → No hay sobreajuste significativo.
  - El MAPE es alto, pero esto es esperable en conjuntos con alta cantidad de ceros o muy bajas ventas (lo cual ocurre aquí).

-  Feature Importance:

Se identificó que variables como `base_price`, `available_quantity`, `is_discounted`, `shipping_is_free`, y `days_since_created` tienen alta influencia en la predicción de ventas.

## 5. Insights para Marketing y Negocio

-  Hipótesis pendientes de validar:
  - Publicaciones con precio base más bajo, envío gratuito y descuento activo tienden a vender más.
  - La antigüedad de la publicación (`days_since_created`) impacta significativamente las ventas. Las ventas iniciales son críticas.
-  Aplicaciones:
  - Optimizar campañas de promoción para productos con alto stock disponible y bajo descuento.
  - Recomendar a vendedores condiciones óptimas de publicación para maximizar ventas (ej. usar envío gratis o establecer un descuento inicial).
-  Casos de uso futuros:

- Recomendación dinámica de precio basada en performance histórica.
- Clustering de productos por potencial de ventas.
- Detección temprana de publicaciones con bajo desempeño esperado.
- Predicción de stock óptimo por vendedor o categoría.