Informe Técnico – Clasificación de Publicaciones MercadoLibre  
“Nuevo” vs “Usado”

*Fecha: 10/07/2025*

# 1. Introducción y Contexto

En el marco del desafío propuesto por Mercado Libre, se desarrolló un sistema de aprendizaje automático capaz de predecir si un artículo listado en el marketplace está en condición \*\*nuevo\*\* o \*\*usado\*\*. El objetivo principal de negocio es minimizar el riesgo de clasificar productos nuevos como usados, manteniendo al mismo tiempo una precisión global superior al 86 %, umbral definido por los requisitos del reto.

# 2. Análisis Exploratorio de Datos (EDA)

Se inspeccionaron 100 000 registros proporcionados en `MLA\_100k\_checked\_v3.jsonlines`. El análisis reveló un moderado desbalance: 58 % de artículos \*\*nuevos\*\* y 42 % \*\*usados\*\*. Se evaluaron distribuciones de precio, cantidad de fotos, modalidades de envío y reputación del vendedor, detectándose valores extremos y ausencias (<5 % de celdas). Las variables \*\*price\*\*, \*\*initial\_quantity\*\* y \*\*sold\_quantity\*\* mostraron asimetría positiva; se aplicó log‑transformación para estabilizar la varianza.

# 3. Selección e Ingeniería de Características

Tomando como referencia los notebooks \*01\_eda\_feature\_selection\* y \*02\_feature\_engineering\_pipeline\*, se diseñó un pipeline reproducible que genera \*\*44 características\*\* a partir de las 56 crudas, entre ellas:  
• \*\*Precio logarítmico\*\* (`log\_price`): mitiga la dispersión en precios y captura escalas relativas.  
• \*\*Volumen histórico del vendedor\*\* (`seller\_volume`): proxia experiencia y reputación.  
• \*\*Flag de garantía\*\* (`has\_warranty`): presencia de política de garantía, más común en ítems nuevos.  
• \*\*Cantidad de imágenes\*\* (`image\_count`): los vendedores de productos nuevos suelen subir más fotos.  
• \*\*Descomposición temporal de `start\_time`\*\* (hora, día de la semana, mes): patrones de alta en stock.  
• \*\*Modalidad de envío\*\* (`shipping\_mode\_express`, `free\_shipping`): ofertas de envío gratuito se asocian a nuevos.  
• \*\*Tipo de publicación\*\* (`listing\_type\_id` one‑hot): distintos planes de publicación reflejan intención comercial.  
• \*\*Conteo de atributos y tags\*\*: mayor riqueza descriptiva suele indicar artículos nuevos.

# 4. Experimentos de Modelado

Se compararon varios algoritmos usando una validación estratificada 80/20:  
• Regresión Logística  
• Random Forest  
• LightGBM  
• Multi‑Layer Perceptron  
• \*\*XGBoost\*\* (ganador)  
  
XGBoost logró la mejor generalización con \*accuracy\* de 87 % y ROC‑AUC de 0.95 antes del ajuste fino de hiper‑parámetros.

# 5. Optimización de XGBoost y Métrica Secundaria

\*\*Método de búsqueda\*\*: se utilizó \*\*`RandomizedSearchCV`\*\* con validación cruzada estratificada, explorando ~90 combinaciones de hiper‑parámetros clave (`max\_depth`, `learning\_rate`, `n\_estimators`, `gamma`, `reg\_alpha`, `reg\_lambda`, `colsample\_bytree`, `min\_child\_weight`, `scale\_pos\_weight`). Esta estrategia ofrece un equilibrio entre cobertura del espacio de búsqueda y costo computacional, y es completamente reproducible (`random\_state=42`).  
  
\*\*Métrica optimizada\*\*: se eligió \*\*ROC‑AUC\*\* porque:  
1. \*\*Robustez ante desbalance\*\* – mide la capacidad discriminatoria sin depender del umbral, lo que evita sesgos debidos a la proporción 58 / 42.  
2. \*\*Visión global\*\* – complementa el \*accuracy\*, que puede sobre‑estimarse si la clase mayoritaria domina.  
  
\*\*Mejores hiper‑parámetros\*\* obtenidos:  
• `max\_depth`: 9 • `learning\_rate`: 0.115 • `n\_estimators`: 287  
• `gamma`: 0.069 • `reg\_alpha`: 1.10 • `reg\_lambda`: 1.37 • `colsample\_bytree`: 0.70  
  
\*\*Ajuste de umbral\*\*: se trasladó el punto de corte a \*\*0.3196\*\*, logrando un \*\*recall del 95 %\*\* para la clase \*nuevo\*, alineado con la prioridad de negocio de no perder ventas potenciales.

# 6. Principales Características Contribuyentes

El análisis de importancia (ganancia en XGBoost y SHAP) resaltó:  
• `log\_price`: 18 % de ganancia. Los productos nuevos muestran rangos de precio diferenciados.  
• `has\_warranty`: 12 %. La presencia de garantía es un fuerte indicio de novedad.  
• `seller\_volume`: 10 %. Vendedores con alto volumen tienden a comercializar ítems nuevos.  
• `image\_count`: 8 %. Más imágenes correlaciona con productos nuevos.  
• `shipping\_mode\_express` y `free\_shipping`: combinados 7 %.  
• `listing\_type\_id\_gold\_premium`: 6 %. Planes pagos premium se usan más para artículos nuevos.

# 7. Resultados Obtenidos

En el conjunto de prueba (20 000 ítems) se obtuvieron:  
• \*\*Accuracy\*\*: 87.3 %  
• \*\*ROC‑AUC\*\*: 0.952  
• \*\*Precision (nuevo)\*\*: 83.7 %  
• \*\*Recall (nuevo)\*\*: 95.0 %  
• \*\*F1‑Score\*\*: 0.890  
  
Se supera así el requisito mínimo de precisión (86 %) y se reduce a solo 5 % el riesgo de etiquetar productos nuevos como usados.

# 8. Conclusiones y Próximos Pasos

El modelo \*\*XGBoost\*\* con un pipeline de 44 características brinda un equilibrio óptimo entre rendimiento predictivo y eficiencia. El enfoque de umbral bajo privilegia el \*recall\* de la clase \*nuevo\*, lo cual es crítico para la estrategia comercial de Mercado Libre.  
  
Próximos pasos:  
• Monitorear \*drift\* y re‑entrenar mensualmente.  
• Implementar explicabilidad item‑a‑item vía SHAP para retroalimentar a vendedores.  
• Explorar \*active learning\* para mejorar la etiqueta \*usado\* en bordes de decisión.