Informe Final - Clasificación de Productos "New vs Used" en Mercado Libre

1. Descripción del Problema

Se desarrolló un modelo de Machine Learning para predecir si un producto listado en el Marketplace de Mercado Libre está en condición "new" o "used", utilizando un dataset de 100k registros. El objetivo fue alcanzar un accuracy mínimo de 0.86, junto con la elección y justificación de una secondary metric.

---

2. Proceso de Desarrollo

2.1 Selección de Variables (Feature Selection)

Se realizó un análisis detallado de las variables disponibles, aplicando los siguientes criterios:

- **Eliminación de variables nulas, redundantes o irrelevantes** (como URLs, imágenes y campos constantes).
- **Variables seleccionadas:**
    - condition (Target)
    - base_price, price, original_price
    - buying_mode, listing_type_id
    - sold_quantity
    - title (vectorizado con TF-IDF)
    - category_id (frecuencia relativa)
    - is_official_store (flag binaria)
    - perfil_vendedor (clasificación del seller por perfil de publicaciones)

Variables como seller_id fueron transformadas en perfiles de vendedor (vende solo "new", "used" o mixto). El campo title fue vectorizado con TF-IDF debido a su potencial predictivo.

2.2 Transformaciones Previas

- Se normalizaron variables numéricas (price, base_price, sold_quantity) utilizando MinMaxScaler.
- Variables booleanas como is_official_store fueron transformadas a 0/1.

- Se detectaron y codificaron patrones en texto (e.g., presencia de la palabra “usado” en el título).
- Se aplicó One-Hot Encoding a variables categóricas como buying_mode y listing_type_id.

2.3 Exploratory Data Analysis (EDA)

Se realizó un EDA profundo con foco en:

- **Distribución de la variable target**: Leve desbalance (53% "new" vs 47% "used").
- **Relaciones bivariadas clave**:
    - is_official_store: Alta precisión para identificar productos "new" pero bajo recall.
    - listing_type_id: Categorías premium tienen correlación fuerte con productos nuevos.
    - sold_quantity: Dispersión amplia en productos nuevos, más concentrado en usados.
    - perfil_vendedor: Vendedores "puros" (que venden solo "new" o "used") son buenos predictores, mientras que los mixtos requieren de otras features complementarias.
    - Variables de precio mostraron patrones coherentes con la lógica de mercado (productos nuevos en rangos más amplios y altos).

---

3. Modelado y Evaluación

3.1 Modelo Utilizado

Se seleccionó **XGBoostClassifier** como modelo base por su robustez ante outliers, buen rendimiento en datos tabulares y su capacidad para manejar sparsity (distribuciones dispersas de TF-IDF).

3.2 Métricas de Evaluación

- **Primary Metric**: Accuracy (objetivo ≥ 0.86)
- **Secondary Metric**: F1-Score (Weighted)
    - Justificación: La F1 ponderada permite balancear precisión y recall en un contexto con clases relativamente equilibradas, penalizando desbalances en predicciones erróneas.

3.3 Resultados Obtenidos

- **Accuracy en Test Set**: 0.8681
- **F1-Score Weighted**: 0.8678
- **Matriz de Confusión**:
    - Alta capacidad de predicción en productos nuevos (Precision = 0.86, Recall = 0.90).
    - Performance aceptable en productos usados (Precision = 0.88, Recall = 0.83).
- **Feature Importance**:
    - listing_type_free, buying_mode_buy_it_now, sold_quantity y las primeras dimensiones de TF-IDF del título fueron las features más determinantes.
    - Variables como is_official_store y original_price funcionaron como indicadores de alta precisión, pero de baja cobertura (bajo recall).

---

4. Consideraciones Finales

- Se utilizó como dataset principal el conjunto de entrenamiento provisto (90k registros) para todo el desarrollo del modelo, dado el acotado tiempo de entrega.
- Las transformaciones aplicadas son reproducibles y fácilmente aplicables sobre cualquier conjunto de test adicional.
- La solución es escalable y puede extenderse para incluir nuevas features o ajuste fino de hiperparámetros si fuera necesario.
- La estructura del código y la documentación permiten su fácil interpretación y reutilización en entornos colaborativos.
