Opportunities@MeLi
Code Exercise

Data Scientist

Description:

In the context of Mercadolibre's Marketplace an algorithm is needed to predict if an item listed in the marketplace is new or used.

Your tasks involve the data analysis, designing, processing and modeling of a machine learning solution to predict if an item is new or used and then evaluate the model over held-out test data.

To assist in that task a dataset is provided in 'MLA_100k_checked_v3.jsonlines' and a function 'build_dataset' to read that dataset in 'new_or_used.py'.

For the evaluation, you will use the accuracy metric in order to get a result of 0.86 as minimum. Additionally, you will have to choose an appropriate secondary metric and also elaborate an argument on why that metric was chosen.

Deliverables

The following items should be addressed in your solution keeping in mind collaborative development such as code reproducibility and legibility.

- The scripts or runnable notebooks should include the necessary code needed to define and evaluate a model.
- A document with an explanation on the criteria applied to choose the features, the proposed secondary metric and the performance achieved on that metrics. Optionally, you can deliver an EDA analysis with other format like .ipynb

Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025

Reporte

▸ Repositorio Github

El dataset está compuesto por 100k items disponibles en la tienda de MercadoLibre entre los años 2013 y 2015. Se realizó un *split* para generar un conjunto de testeo del 10%.

Se compone de 45 *features* con las siguientes características:

- **IDs:** 'id', 'parent_item_id', 'official_store_id', 'category_id', 'site_id', 'video_id', 'catalog_product_id', 'deal_ids', 'seller_id', 'descriptions'.
- **Price-related variables:** 'price', 'original_price', 'base_price'.
- **Transaction-related variables:** 'currency_id', 'accepts_mercadopago', 'warranty'.
- **Quantity-related variables:** 'initial_quantity', 'sold_quantity', 'available_quantity'.
- **Post-related variables:** 'title', 'thumbnail', 'pictures', 'permalink', 'secure_thumbnail', 'status', 'sub_status', 'buying_mode', 'condition', 'automatic_relist'.
- **Time-related variables:** 'start_time', 'stop_time', 'date_created', 'last_updated'.
- **Seller-related variables:** 'seller_address', 'listing_type_id'.
- **Tags:** 'tags', 'variations', 'shipping', 'attributes', 'non_mercado_pago_payment_methods'.

También se incluía en el conjunto de entrenamiento la variable 'condition', que fue eliminada por tratarse de la variable objetivo (target). La misma está compuesta por etiquetas 'new' y 'used' con distribución balanceada.

Feature Engineering

Valores faltantes

Se detectaron variables con gran cantidad de valores nulos ('video_id', 'official_store_id', 'original_price', 'catalog_product_id') –menos del 2% de los items contenía información– o completamente vacías ('differential_pricing', 'subtitle', 'coverage_areas','listing_source'). Estas columnas fueron descartadas.

Además, para 'parent_item_id' un 23% eran nulos y para 'warranty' un 60.8%. En este último caso, se observó el contenido de los campos y se encontró gran disparidad en la información contenida (texto descriptivo escrito por el usuario-vendedor con relación a la garantía ofrecida). Se decidió no utilizarla debido a la complejidad y arbitrariedad de los inputs. Podría extraerse valor analizando con un modelo LLM para generar una o varias variables de mayor calidad, pero se optó por no incluirlo en este *sprint* por la demanda de recursos y la poca certeza de que aportara un valor significativo.

Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025

Eliminación de items

Se detectó que la variable ‘sub_status’ incluía una etiqueta para todos aquellos items que no están disponibles en la tienda (“suspendidos”, “expirados” o “eliminados”). Se eliminaron entonces todas aquellas entradas con estas etiquetas, removiendo 891 items (-0.99%) para entrenamiento.

Se observó que algunos items tenían precios irreales (probablemente sean pruebas, no productos en venta) por lo que es estableció una cota superior, quitando aquellos valores por encima de 6.5 millones de pesos (~650k USD en 2015). En total se removieron 10 items.

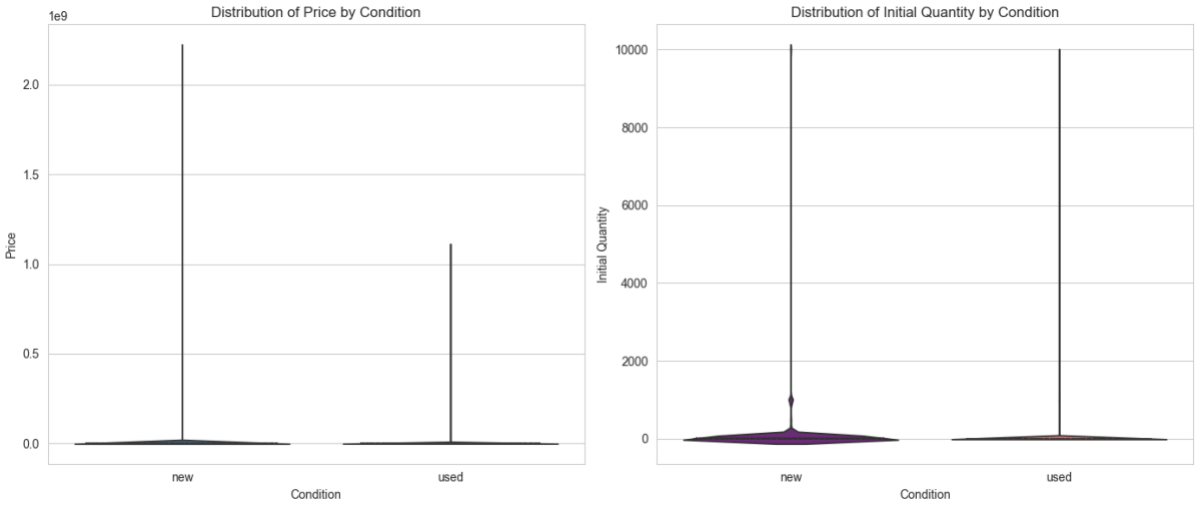
Eliminación de *features*

Como se mencionó anteriormente se eliminaron aquellas columnas vacías o con elevado número de valores nulos. También se removieron las que incluían datos repetidos o similares, dejando sólo la más representativa, por ejemplo, sólo se dejó ‘price’ y se quitaron ‘original_price’ y ‘base_price’.

Se descartaron también las variables de IDs, por tratarse de variables independientes, y las relacionadas con la metadata de la publicación en la web. No fueron tenidas en consideración algunos diccionarios de etiquetas con valores repetidos, desconocidos o dispersos: 'variations', 'attributes' y 'non_mercado_pago_payment_methods'.

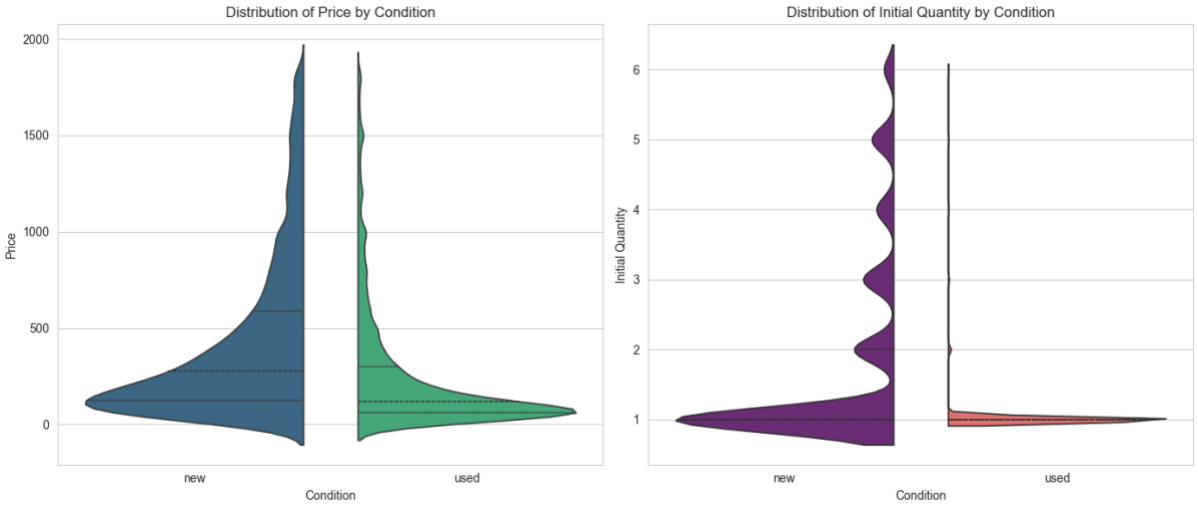
Variables continuas

En las variables ‘price’ y ‘initial_quantity’ se observó una distribución de cola larga altamente sesgada hacia los valores inferiores, por tal motivo, fue necesario aplicar un tratamiento de *outliers* y escalado.



Se realizó un escalado por rango intercuartílico (IQR) con factor 1.5 y luego escalado Min-Max para acomodar los valores en el rango 0 a 1, al igual que las demás variables.

Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025



Variables discretas

Se transformaron las variables binarias para ser representadas como booleanos (‘currency_id’ == ‘USD’ → **‘is_USD’** == True).

Variables categóricas

Se aplicó One-hot Encoding a **‘listing_type_id’** para generar variables dummy, para categorizar el tipo de cliente-vendedor. De igual modo se hizo para ‘buying_mode’ y ‘status’.

Posteriormente, se observó que los coeficientes del modelo le daban demasiada importancia a ‘initial_quantity_scaled’, por lo que se decidió reemplazarla por variables categóricas (**'quant_single_unit', 'quant_small, ‘quant_large'**) diferenciando entre cantidades de 1, 10 y superiores respectivamente.

Variables sintéticas

- Se procesó la feature ‘date_created’ (string) para detectar la hora de publicación (‘time_created’) y el día de semana (‘day_of_week’). Luego, se decidió convertirlas en categóricas **‘is_weekend’** y **‘is_working_hours’**.
- **'parent_item_id' → 'has_parent_item':** *True* si existe una referencia a un item padre.
- **'official_store_id' → 'has_store':** *True* si existe una referencia a una tienda oficial.
- **'price' → 'high_ticket':** Items de precio alto, por encima del percentil 75 (> 80 usd).
- **'stop_time' | 'start_time' → 'duration':** Duración de la publicación.
- **Diccionario ‘tag’ → 'good_quality_thumbnail' / 'dragged_bids_and_visits' / 'dragged_visits' / 'poor_quality_thumbnail' / 'free_relist' (categóricas)**
- **Diccionario 'shipping' → 'local_pick_up' / 'free_shipping' (categóricas)**

Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025

Criterio de selección de *features*

De las 38 features iniciales se seleccionaron 5 y se generaron 10 nuevas, entrenando el primer modelo con 15 parámetros seleccionados.

Selección manual

La selección inicial se basó en lo observado en el análisis inicial (EDA) y la correlación entre las features y la variable target. El criterio de selección se basó en aquellos parámetros que eran mejores predictores de alguna de las dos clases de target.

Se identificó como mejores predictortes a la categoría ‘listing_type_id_free’ para la clase ‘used’ y, por el contrario, las features ‘automatic_relist’ y ‘listing_type_id_silver’ para la clase ‘new’. Se observa además que varias de las categorías de clientes de pago se correlacionan con la venta de productos nuevos. Se decidió agruparlos bajo un solo parámetro ‘free_tier’ y remover las variables dummy.

Luego del primer entrenamiento se descartaron las categóricas derivadas de ‘status’ ya que no estaban contribuyendo a mejorar la predicción. Se redujo casi a la mitad el número de parámetros sin deterioro en la performance.

A continuación, se exploró la contribución de distintas variables sintéticas.

Chi-cuadrado + RFE

Se combinó Chi-cuadrado con Automatic Feature Selection (RFE) para permitir al modelo seleccionar una combinación de features óptima de forma automática y compararla con la selección manual realizada previamente.

Como resultado se observó que los parámetros seleccionados en su gran mayoría era los mismos, aunque se le daba mucha importancia a ‘duration’ y ‘automatic_relist’, por lo que se decidió incluirlas en la selección de features final.

Selección final

'accepts_mercadopago', 'automatic_relist', 'price_scaled', 'free_tier',

'buying_mode_buy_it_now', 'buying_mode_classified', 'duration',

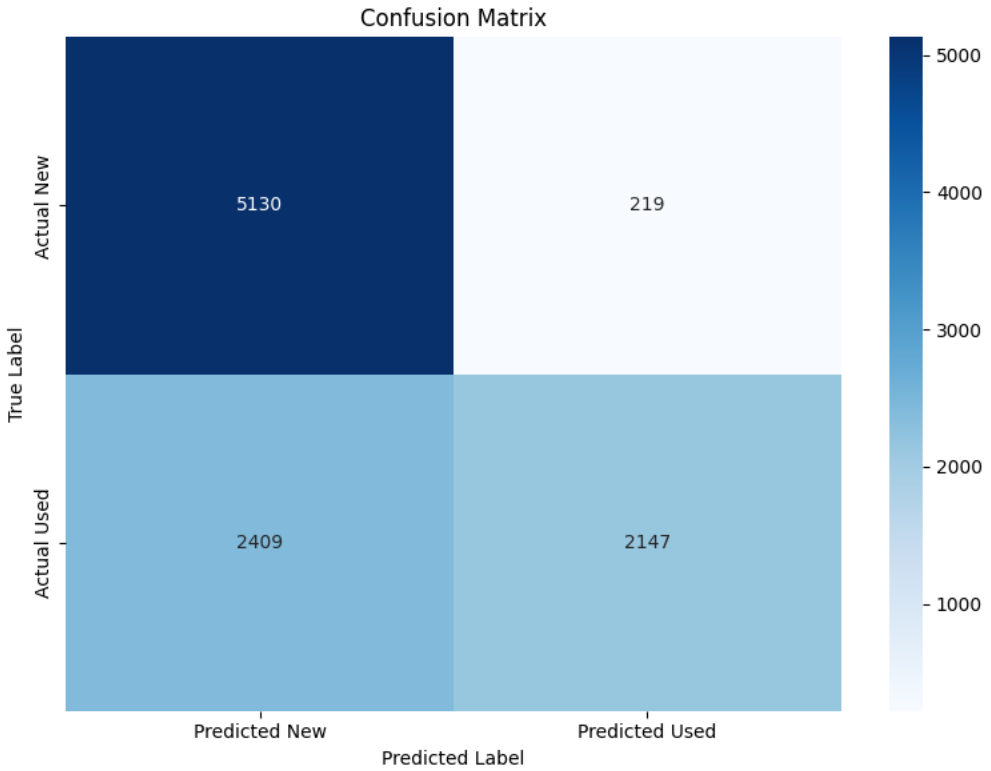
'quant_single_unit', 'has_parent_item', 'tag_good_quality_thumbnail',

'tag_free_relist', 'tag_dragged_visits', 'tag_dragged_bids_and_visits'

Métrica secundaria propuesta

Desde la primera iteración de entrenamiento de modelos, se detectó que el modelo lograba detectar bien la clase ‘new’ (sin FP), pero no lograba performar para la clase ‘used’ (muchos FN), sin ser mejor que una selección aleatoria para dicha clase.

Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025



Dado el marcado desequilibrio entre el desempeño por clases, con un marcado Recall, se optó por utilizar F1-score como métrica secundaria, para poder observar mejor el balance entre Precision y Recall. Al ser la media armónica de estas dos métricas, F1-score penaliza el desbalance entre clases.

Performance alcanzado

Best Logistic Regressor

Se comenzó entrenando un modelo de regresión logística e iterando en la selección de features. Se observó que al incorporar regularización L2 se logra una leve mejora e empareja la influencia de cada *feature.*

- **Accuracy:** 0.7990
- **F1-score:** 0.7950

Best Random Forest

Con la incorporación del modelo Random Forest se alcanzó una performance más equilibrada y mejor capacidad de clasificación, especialmente al combinarlo con un ajuste de hiperparámetros realizado por búsqueda aleatoria (RandomizedSearchCV) utilizando 'accuracy' como métrica de referencia.

- **Accuracy:** 0.8298
- **F1-score:** 0.8343

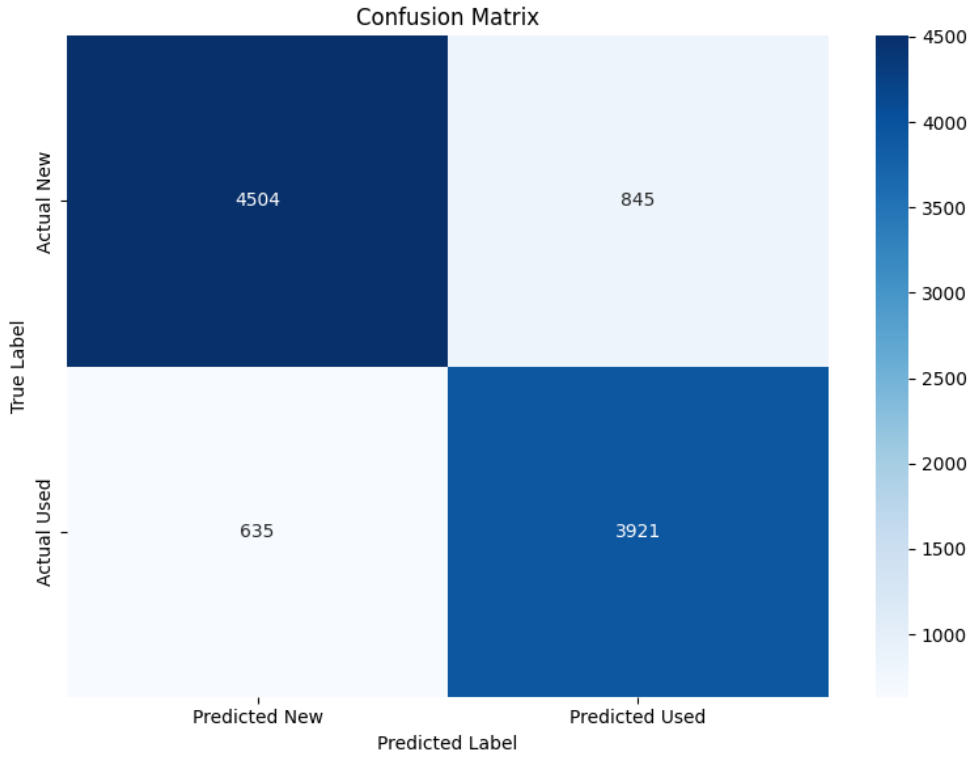
Si bien el accuracy aún se encuentra lejos del valor buscado, se logró una mejora significativa en el balance entre clases, incrementado F1-score.

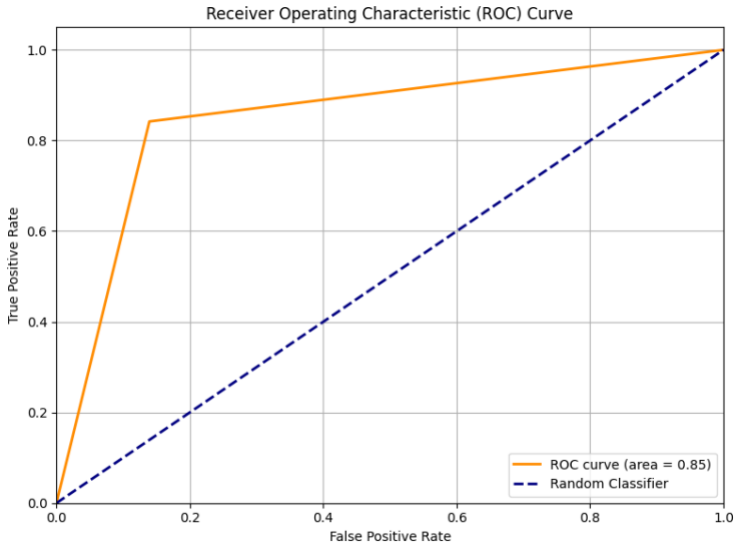
Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025

Best XGBoost

Finalmente, se implementó XGBoost logrando el mejor *trade-off* entre Precision y Recall. Los mejores resultados se obtuvieron combinando la selección manual de *features* junto con las conclusiones extraídas con la técnica RFE. Se aplicó además ajuste de hiperparámetros para maximizar 'accuracy', alcanzando los siguientes resultados:

- **Accuracy:** 0.8503
- **F1-score:** 0.8343





Autor: Alejandro Lloveras
Fecha: 21 de Julio de 2025

Appendix

Dataset schema


```
{
  "seller_address": {
    "country": {
      "name": "Argentina",
      "id": "AR"
    },
    "state": {
      "name": "Capital Federal",
      "id": "AR-C"
    },
    "city": {
      "name": "San Cristóbal",
      "id": "TUxBQlNBTjkwNTZa"
    }
  },
  "warranty": null,
  "sub_status": [],
  "condition": "new",
  "deal_ids": [],
  "base_price": 80.0,
  "shipping": {
    "local_pick_up": true,
    "methods": [],
    "tags": [],
    "free_shipping": false,
    "mode": "not_specified",
    "dimensions": null
  },
  "non_mercado_pago_payment_methods": [
    {
      "description": "Transferencia bancaria",
      "id": "MLATB",
      "type": "G"
    },
    {
      "description": "Acordar con el comprador",
      "id": "MLAWC",
      "type": "G"
    },
    {
      "description": "Efectivo",
      "id": "MLAMO",
      "type": "G"
```

```
    }
  ],
  "seller_id": 8208882349,
  "variations": [],
  "site_id": "MLA",
  "listing_type_id": "bronze",
  "price": 80.0,
  "attributes": [],
  "buying_mode": "buy_it_now",
  "tags": [
    "dragged_bids_and_visits"
  ],
  "listing_source": "",
  "parent_item_id": "MLA6553902747",
  "coverage_areas": [],
  "category_id": "MLA126406",
  "descriptions": [
    "{'id': 'MLA4695330653-912855983'}"
  ],
  "last_updated": "2015-09-05T20:42:58.000Z",
  "international_delivery_mode": "none",
  "pictures": [
    {
      "size": "500x375",
      "secure_url": "https://a248.e.akamai.net/mla-s1-
p.mlstatic.com/5386-MLA4695330653_052013-O.jpg",
      "max_size": "1200x900",
      "url": "http://mla-s1-p.mlstatic.com/5386-MLA4695330653_052013-
O.jpg",
      "quality": "",
      "id": "5386-MLA4695330653_052013"
    },
    {
      "size": "500x375",
      "secure_url": "https://a248.e.akamai.net/mla-s1-
p.mlstatic.com/5361-MLA4695330653_052013-O.jpg",
      "max_size": "1200x900",
      "url": "http://mla-s1-p.mlstatic.com/5361-MLA4695330653_052013-
O.jpg",
      "quality": "",
      "id": "5361-MLA4695330653_052013"
    }
  ],
  "id": "MLA4695330653",
  "official_store_id": null,
  "differential_pricing": null,
```

```
  "accepts_mercadopago": true,
  "original_price": null,
  "currency_id": "ARS",
  "thumbnail": "http://mla-s1-p.mlstatic.com/5386-MLA4695330653_052013-
I.jpg",
  "title": "Auriculares Samsung Originales Manos Libres Cable Usb
Oferta",
  "automatic_relist": false,
  "date_created": "2015-09-05T20:42:53.000Z",
  "secure_thumbnail": "https://a248.e.akamai.net/mla-s1-
p.mlstatic.com/5386-MLA4695330653_052013-I.jpg",
  "stop_time": 1446669773000,
  "status": "active",
  "video_id": null,
  "catalog_product_id": null,
  "subtitle": null,
  "initial_quantity": 1,
  "start_time": 1441485773000,
  "permalink": "http://articulo.mercadolibre.com.ar/MLA4695330653-
auriculares-samsung-originales-manos-libres-cable-usb-oferta-_JM",
  "sold_quantity": 0,
  "available_quantity": 1
}
```