Reporte Prueba Técnica: Data Scientist .

Lina Marcela Aguilar Gonzalez

June 14, 2025

Resumen

El propósito de esta prueba técnica fue evaluar mis habilidades en Ciencia de Datos y Machine Learning aplicadas al contexto del marketplace de Mercado Libre. Se trabajó con un dataset de 100.000 publicaciones de productos, incluyendo características detalladas del ítem, del vendedor y del comportamiento de ventas.

Mi enfoque para resolver el desafío se centró en dos objetivos predictivos clave:

1. **Predicción del precio final de los productos** mediante modelos de regresión, para estimar con mayor precisión el valor de mercado de un ítem dado.
2. **Predicción de la probabilidad de venta** (clasificación binaria), a partir de atributos del producto y del vendedor, con el fin de identificar qué características favorecen la conversión de una publicación.

Para ello, estructuré la solución en clases modulares que siguen buenas prácticas de programación orientada a objetos:

- **DataAnalyzer:** análisis exploratorio de datos y calidad del dataset.
- **FeatureEngineer:** creación y selección de variables relevantes.
- **SalesPredictor y PricePredictor:** especializadas en las tareas de clasificación y regresión respectivamente.

Además, se extrajeron insights accionables para equipos de negocio y marketing que permitirían mejorar la estrategia de publicación y posicionamiento de productos en la plataforma.

1 Análisis Exploratorio de Datos (EDA)

Para comenzar el desarrollo del proyecto, se realizó un análisis exploratorio del dataset proporcionado, el cual contiene información de 100.000 publicaciones del marketplace de Mercado Libre. El análisis tuvo como objetivos principales:

- Evaluar la calidad y distribución de las variables clave.
- Identificar valores atípicos (outliers) y sesgos en los datos.
- Establecer criterios de limpieza y filtrado para el modelado.

1.1 Variables de interés: `sold_quantity` y `price`

Se prestó especial atención a dos variables fundamentales para los objetivos de predicción: la cantidad de productos vendidos (`sold_quantity`) y el precio final de los productos (`price`). A continuación se presentan los estadísticos descriptivos de ambas:

Distribución de `sold_quantity`:

count	96041.000000
mean	2.485595
std	43.549101
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	8676.000000

Se observa una fuerte concentración de valores en cero: aproximadamente el 82 % de las publicaciones no registraron ninguna venta. Esto implica un dataset altamente desbalanceado, lo cual representa un desafío importante para modelos de clasificación. En caso de predecir si un producto se venderá o no, este desbalance debe ser considerado explícitamente en la estrategia de modelado.

Distribución de `price`:

```
count    96041.000000
mean       761.437883
std       1418.625975
min          0.840000
25%         89.800000
50%        229.000000
75%        690.000000
max      10000.000000
```

La variable `price` presenta también una distribución fuertemente sesgada hacia valores bajos, y contiene numerosos valores atípicos, especialmente en el rango alto de precios. Dado que el precio será la variable objetivo en una de las tareas de regresión, estos outliers fueron analizados y tratados con cuidado en las etapas posteriores.

1.2 Filtrado geográfico

Como parte de la limpieza inicial, se decidió restringir el análisis y modelado a los productos ofrecidos exclusivamente en **Argentina**, con el fin de reducir heterogeneidades relacionadas con mercados locales, condiciones económicas, y políticas comerciales entre países. Esta decisión también favorece la consistencia del modelo y mejora la calidad de los resultados.

1.3 Visualización

Se graficaron histogramas para todas las variables numéricas, lo cual permitió detectar patrones generales de distribución y confirmar la presencia de sesgos severos en `price` y `sold_quantity`. Estas visualizaciones se incluyen en el apéndice del presente informe.

1.4 Revisión general del dataset

El dataset cuenta con **99.989 registros** y **26 columnas**, con una cobertura bastante completa de valores no nulos. A continuación, se destacan algunos hallazgos relevantes del resumen estructural:

- La mayoría de las variables presentan valores completos, con excepción de:
  - `sub_status`, con solo 986 valores no nulos (~1 %).
  - `warranty`, con un 61 % de valores faltantes.
- Las variables numéricas como `base_price` y `price` presentan valores máximos extremadamente altos (>2.200 millones), lo cual confirma la presencia de outliers severos.
- En las variables categóricas:
  - `tags` muestra una distribución muy desbalanceada, con más del 72 % de registros concentrados en un único valor.
  - `seller_country` contiene únicamente datos de Argentina, por el filtrado hecho previamente.

Estas observaciones se tuvieron en cuenta en las etapas de limpieza y generación de variables. Las columnas con alta proporción de valores nulos fueron descartadas o transformadas según su relevancia para el análisis.

2 Feature Engineering

En esta etapa se realizó un proceso riguroso de limpieza, transformación y creación de variables con el objetivo de mejorar la calidad de los datos y facilitar el aprendizaje de los modelos predictivos. Las decisiones se fundamentaron tanto en criterios estadísticos como en el conocimiento del dominio del negocio.

2.1 Observaciones y decisiones generales

Eliminación de columnas

Se eliminaron columnas que presentaban un alto porcentaje de valores faltantes o que no aportaban información relevante para las tareas de predicción:

- `sub_status`: contenía más del 99 % de valores faltantes.
- `attributes` y `variations`: con 88 % y 92 % de valores faltantes respectivamente, sin evidencia de relación con el precio de venta.
- `seller_country`: dado que el dataset contenía exclusivamente publicaciones de Argentina, esta variable no aportaba variabilidad. En su lugar se conservó `seller_province`, con un número manejable de categorías, y se eliminó `seller_city`, debido a su alta cardinalidad.
- `warranty`: presentaba un 61 % de valores faltantes y más de 10,000 valores únicos, lo que sugiere que era un campo de entrada libre difícil de estandarizar.

Otras decisiones importantes

- Se evaluó la relación entre las variables `base_price` y `price`, y al verificar que en la mayoría de los casos eran iguales, se eligió utilizar una como variable objetivo y se descartó la otra como predictor.

- De la variable `pictures`, se extrajo la cantidad de imágenes por publicación, convirtiéndola en una nueva variable numérica.

2.2 Transformaciones y creación de variables

- Se unificaron y estandarizaron variables booleanas originalmente representadas como categóricas, tales como `shipping_admits_pickup`, `shipping_is_free` e `is_new`, para facilitar su interpretación por los modelos.

- Se crearon nuevas variables a partir del título del producto: la longitud del título en caracteres (`title_len`) y el número de palabras (`title_word_count`), con el fin de capturar posibles señales semánticas o de calidad de publicación.

- A partir de la fecha de creación, transformada al formato `datetime`, se generaron variables temporales como el año, mes y día de publicación, así como los días transcurridos desde la creación.

2.3 Codificación de variables categóricas

Para preparar las variables categóricas para su uso en modelos de aprendizaje automático:

- Se aplicó codificación ordinal a la variable `seller_id`, asignando a cada vendedor un identificador numérico único.

- Se utilizó codificación one-hot para variables categóricas sin orden inherente, como `seller_province`, `seller_loyalty`, `buying_mode`, `shipping_mode` y `status`.

- En la columna `tags`, que contiene listas de etiquetas asociadas a cada publicación, se realizó una transformación en varios pasos: primero, se convirtieron los textos en listas interpretables; luego, se binarizaron las etiquetas, creando una columna por cada etiqueta posible e indicando su presencia o ausencia por publicación.

Este proceso de ingeniería de características permitió reducir la dimensionalidad, mejorar la representatividad de los datos y facilitar el entrenamiento de modelos tanto de clasificación como de regresión.

3 Modelo Predictivo

A partir del dataset entregado, se identificaron múltiples posibles enfoques de modelado:

- **Regresión**: para predecir la cantidad vendida o el precio final de los productos.

- **Clasificación**: para predecir si un producto se venderá o no, formulando el problema como una variable binaria (`sold_quantity >0`).

- **Clustering**: con el fin de agrupar productos similares y realizar segmentación de mercado.

- **Detección de anomalías**: para identificar publicaciones con comportamientos atípicos.

En este trabajo, se abordaron dos tareas principales: la predicción del **precio final del producto** mediante modelos de regresión, y la predicción de la **probabilidad de venta** de una publicación mediante clasificación binaria.

3.1 Regresión: Predicción del precio

Inicialmente se evaluó la capacidad de modelos de regresión para predecir directamente la variable `price`. Se probaron modelos como **Random Forest Regressor** y **XGBoost Regressor**, con resultados iniciales poco satisfactorios:

- **Random Forest**:
  - MAE: 60,116.98
  - RMSE: 7,857,315.82
  - R^2: 0.0001

- **XGBoost**:
  - MAE: 60,318.33
  - RMSE: 7,857,283.15
  - R^2: 0.0001

Debido a estos resultados, se decidió transformar la variable objetivo aplicando logaritmo natural (log-price) y eliminar valores atípicos extremos, aunque esto implica un riesgo si el modelo se usara posteriormente en producción. Con estas modificaciones, se obtuvieron métricas considerablemente mejores:

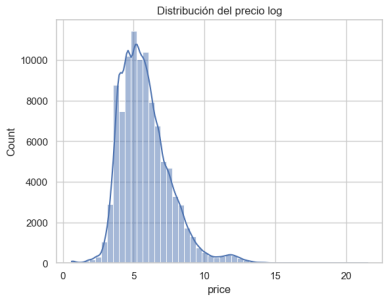

Figura 1: Distribución de la variable `log(price)` después de la transformación

- **Validación cruzada (5 folds)**:
  - Random Forest: RMSE medio = 1.11 (± 0.00)
  - XGBoost: RMSE medio = 1.16 (± 0.00)
- **Evaluación final**:
  - **Random Forest**:
    - MAE: 731.48
    - RMSE: 2,065.24
    - R^2: 0.1158
  - **XGBoost**:
    - MAE: 756.40
    - RMSE: 2,122.83

A pesar de la mejora, los modelos continúan presentando dificultades para capturar con precisión el comportamiento del precio. Se construyó una línea base (**baseline**) calculando el MAE al predecir simplemente la media del precio, y se comprobó que los modelos efectivamente aprenden, reduciendo el error a casi la mitad. Sin embargo, la **alta varianza en los precios** y la presencia de **outliers extremos** siguen afectando negativamente el RMSE y el R^2.

Análisis de correlación y relevancia de variables

Se construyó un mapa de correlación entre las variables numéricas y el logaritmo del precio (`log(price)`), el cual se incluye en el apéndice. También se evaluó la importancia de las variables utilizando las métricas internas de los modelos de árbol (feature importance). Aunque estas visualizaciones fueron útiles para el análisis exploratorio, no condujeron a mejoras significativas en el rendimiento de los modelos. La gráfica de importancia de variables más relevantes se presenta en la Figura 6.

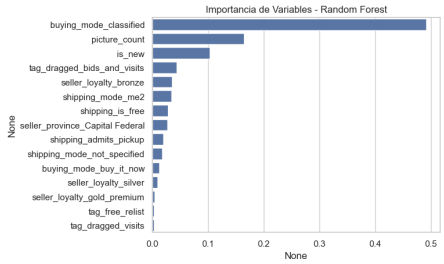

Figura 2: Importancia de variables según modelo de árbol

Conclusión de la regresión

Aunque los modelos mejoraron tras la transformación de la variable objetivo y la eliminación de outliers, los resultados no fueron suficientemente robustos para un modelo de predicción de precio confiable en producción. Se recomienda explorar otros enfoques alternativos, como modelar rangos de precios, usar modelos de series temporales si se cuenta con histórico, o complementar con información externa.

Cambio de enfoque: Clasificación binaria

Dado que los modelos de regresión no lograron predecir el precio con una precisión aceptable, se cambió el enfoque hacia una tarea de clasificación binaria, formulando el problema como la predicción de si un producto se venderá o no (variable `was_sold`, igual a 1 si `sold_quantity` >0).

Este tipo de enfoque permite al negocio priorizar publicaciones con alta probabilidad de conversión. Sin embargo, se encontró una distribución de clases altamente desbalanceada, donde la mayoría de los productos no registraban ventas, lo que representó un reto importante para el modelado.

Para abordar este desafío, se desarrolló una clase modular denominada `SalesPredictor`, que incluye las siguientes etapas:

- **Generación de variable objetivo:** Se creó una variable binaria para indicar si un producto fue vendido.
- **Manejo del desbalanceo:** Se aplicó la técnica SMOTE para balancear la clase minoritaria en el conjunto de entrenamiento, y se utilizaron modelos que permiten ponderar las clases como *Random Forest* con `class_weight='balanced'` y *XGBoost* con `scale_pos_weight`.
- **Entrenamiento y evaluación:** Se entrenaron los modelos y se evaluaron mediante precisión, recall, F1-score y matriz de confusión.

- **Ajuste de umbral de decisión:** Se intentó mejorar el rendimiento ajustando el umbral de decisión en las predicciones probabilísticas, en lugar de utilizar el valor estándar de 0.5.

- **Validación cruzada:** Se realizó validación cruzada estratificada de 5 pliegues, optimizando el threshold en cada iteración.

Los resultados obtenidos con el modelo XGBoost, sin ajustar el umbral, se adjuntan en el apendice:

Interpretación: El modelo presenta un muy buen rendimiento para detectar los productos que no se venden (clase 0), con una precisión del 93 % y un recall del 92 %. Esto significa que comete pocos errores al identificar publicaciones sin conversión.

Para los productos vendidos (clase 1), el desempeño es más moderado:

- **Precisión:** 64 % → De todas las predicciones positivas, solo el 64 % corresponde a productos efectivamente vendidos. Hay un 36 % de falsos positivos.

- **Recall:** 68 % → El modelo logra detectar el 68 % de los productos vendidos, pero omite un 32 %.

- **F1-score:** 0.66 → Refleja un equilibrio moderado entre precisión y recall, pero aún con margen de mejora.

Dado que la clase de productos vendidos es minoritaria, este rendimiento es esperable, aunque puede ser crítico mejorar la detección de oportunidades reales de venta.

Posteriormente, se evaluó la posibilidad de mejorar el modelo ajustando el **umbral de decisión**. A través de validación cruzada con tuning de thresholds, se obtuvieron los siguientes promedios:

- **F1-score promedio:** 0.6493

- **Precisión promedio:** 0.6824

- **Recall promedio:** 0.6193

A pesar del ajuste, los resultados fueron similares a los obtenidos inicialmente, confirmando que el modelo presenta cierta capacidad de aprendizaje, pero que la naturaleza desbalanceada y el comportamiento de los datos limitan su capacidad de generalización. Se recomienda continuar experimentando con técnicas de *feature selection*, *ensemble models* o incluso reformulaciones del problema.

4 Análisis de Resultados

4.1 Evaluación de modelos

Se implementaron dos enfoques predictivos sobre el dataset: uno de regresión para estimar el precio de los productos, y otro de clasificación binaria para determinar si un producto se vende o no.

Predicción del precio (Regresión)

Los modelos de regresión mostraron un rendimiento limitado. Si bien lograron capturar algunas tendencias generales en los precios, la capacidad explicativa fue baja, indicando que los modelos no logran ajustarse adecuadamente a la variabilidad observada.

Esto puede deberse a varios factores:

- La presencia de valores extremos (outliers) que afectan las métricas de error.

- Una representación incompleta o subóptima de variables clave que inciden en el precio real de mercado (como la marca, la condición del producto, o el tipo de categoría).

- La falta de enriquecimiento de atributos textuales o multivalorados como `attributes`, `tags` y `pictures`.

Por tanto, aunque el modelo puede servir como aproximación inicial, se considera que aún tiene un amplio margen de mejora para ser usado en producción o como herramienta de pricing más precisa.

Predicción de ventas (Clasificación)

El modelo de clasificación logró un muy buen desempeño al identificar publicaciones que no se venden, mostrando alta precisión y recall para esa clase. Sin embargo, su rendimiento fue más limitado para detectar productos que sí se venden, con una proporción significativa de falsos positivos y falsos negativos.

Este comportamiento es esperable, dado el fuerte desbalance de clases en el dataset: la mayoría de los productos no registra ventas, lo cual complica la capacidad del modelo para aprender patrones representativos de la clase minoritaria.

Para mitigar este efecto, se ajustó el umbral de decisión y se aplicó validación cruzada enfocada en maximizar el F1-score. Esta estrategia permitió mejorar ligeramente el equilibrio entre precisión y recall para la clase positiva, aunque sin lograr una separación completamente robusta.

4.2 Limitaciones del enfoque actual

- **Desbalance de clases:** La fuerte asimetría en la variable objetivo (venta sí/no) compromete el rendimiento del modelo para identificar oportunidades reales de conversión.
- **Complejidad estructural de las variables:** Algunas columnas clave como `attributes`, `variations` y `pictures` no fueron explotadas completamente por su formato complejo (listas, objetos anidados o textos estructurados), lo que limita la expresividad del modelo.
- **Outliers no tratados a fondo:** La regresión de precios se ve afectada por valores extremos muy alejados del rango típico, lo cual impacta negativamente las métricas de error y la estabilidad de los modelos.
- **Falta de datos contextuales externos:** El precio y la probabilidad de venta pueden depender de factores externos como promociones, estacionalidad, reputación del vendedor o competencia directa, los cuales no están representados en el dataset.

Recomendación: Para futuros desarrollos se recomienda enriquecer los atributos con codificación estructurada o embeddings, y aplicar modelos más robustos a ruido y outliers como LightGBM, CatBoost o redes neuronales con atención.

5 Insights para Marketing y Negocio

A partir del análisis exploratorio y los modelos predictivos desarrollados, se identificaron diversos patrones y hallazgos que pueden ser útiles para los equipos de marketing, comercial y producto de Mercado Libre. A continuación, se destacan los principales insights:

1. Alta proporción de publicaciones sin ventas

El 82 % de las publicaciones analizadas no registran ninguna venta. Esto sugiere una oportunidad significativa para:

- **Optimizar la calidad de las publicaciones:** Redactar mejores títulos, agregar imágenes más atractivas y completar atributos relevantes podría mejorar la conversión.
- **Revisar la estrategia de pricing:** Algunos productos podrían estar sobrevalorados o fuera del rango competitivo del mercado.
- **Segmentar vendedores con baja conversión:** Se podrían ofrecer recomendaciones automáticas a vendedores con tasas de conversión bajas para ayudarles a mejorar.

2. Hipótesis sobre variables relevantes para predecir ventas

Si bien el modelo desarrollado no fue concluyente en la identificación de variables predictoras fuertes para las ventas, se identifican algunas **hipótesis de interés** que podrían ser evaluadas en análisis futuros:

- **Precio moderado:** Es posible que los productos en rangos de precio medios tengan mayor probabilidad de conversión, en comparación con productos muy económicos o extremadamente costosos.
- **Condición del producto:** La variable `is_new`, que indica si un producto es nuevo, podría estar asociada a una mayor intención de compra, aunque esto requiere validación empírica.
- **Condiciones logísticas:** Publicaciones que ofrecen envío gratuito o retiro en domicilio podrían ser más atractivas para los usuarios, y por tanto tener mayor tasa de conversión.

Estas hipótesis podrían ser contrastadas en trabajos posteriores mediante modelos explicativos, análisis de correlaciones o estudios controlados. De comprobarse su validez, podrían ser utilizadas para desarrollar estrategias de optimización de publicaciones, segmentación de campañas y mejoras en la experiencia del usuario.

3. Potencial de un modelo de scoring comercial

Con base en el modelo de clasificación binaria, es viable desarrollar un **score de publicación** que anticipe la probabilidad de venta de un ítem nuevo. Este score podría ser útil para:

- Priorizar productos en los resultados de búsqueda interna.
- Alertar a los vendedores sobre publicaciones con baja probabilidad de éxito.
- Enviar recomendaciones automáticas para mejorar la performance de publicaciones.

4. Casos de uso adicionales recomendados

- **Segmentación de vendedores:** Clasificar vendedores según su tasa de éxito, mix de productos, cumplimiento logístico y engagement.
- **Detección de outliers y anomalías:** Identificar publicaciones con precios anómalos que podrían deberse a errores de carga o estrategias poco competitivas.
- **Predicción de volumen de ventas:** Modelar la cantidad esperada de unidades vendidas para mejorar la planificación de stock y logística.

Conclusión: El análisis realizado permite identificar puntos críticos de optimización en las publicaciones del marketplace. Con ajustes en pricing, logística y presentación de productos, combinados con modelos predictivos, se pueden generar recomendaciones personalizadas y mejoras en la eficiencia comercial de la plataforma.

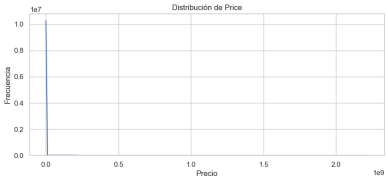

Figura 3: Histograma del precio de los productos (`price`).

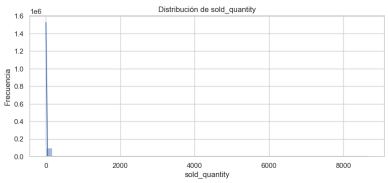

Figura 4: Histograma de la cantidad de ventas (`sold_quantity`).

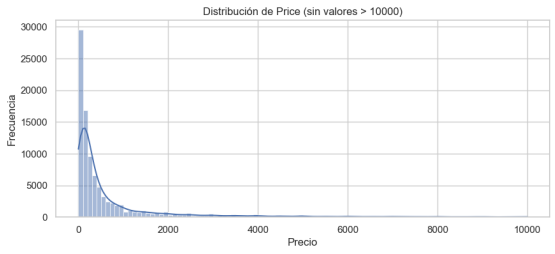

Figura 5: Distribución de precios (`price` $<$ 10,000) para publicaciones en Argentina. Se observa un sesgo hacia valores bajos, con alta concentración por debajo de los 2,000 pesos.

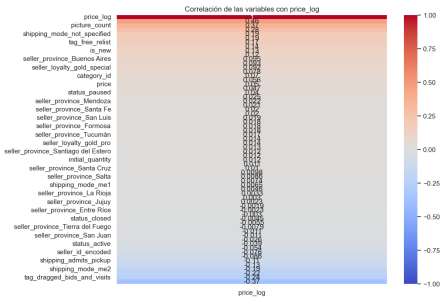

Figura 6: Correlación de las variables con price-log

Cuadro 1: Reporte de clasificación del modelo XGBoost

Clase	Precisión	Recall	F1-score	Soporte
0 (No vendido)	0.93	0.92	0.93	16614
1 (Vendido)	0.64	0.68	0.66	3384
Accuracy	0.88 (sobre 19998 muestras)			
Macro promedio	0.79	0.80	0.79	–
Promedio ponderado	0.88	0.88	0.88	–

Cuadro 2: Matriz de confusión

	Predicho 0	Predicho 1
Real 0	15317	1297
Real 1	1095	2289