Análisis Predictivo - 2025Q2

# Examen 3
# Competencia Kaggle

Catalina Trevisan - 64990

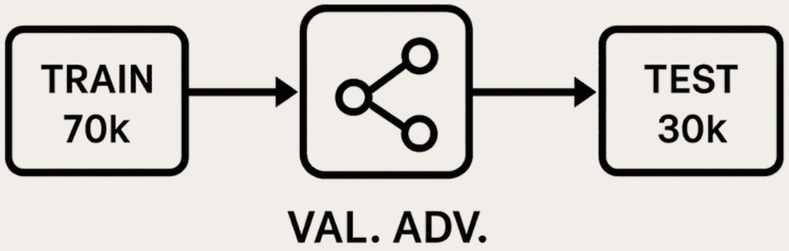
# Introducción

**Objetivo**: predecir si un ´ítem es nuevo o usado.

Train shape: (70.000, 45)
Test shape: (30.000, 44)

Paso inicial: **Validación adversarial**

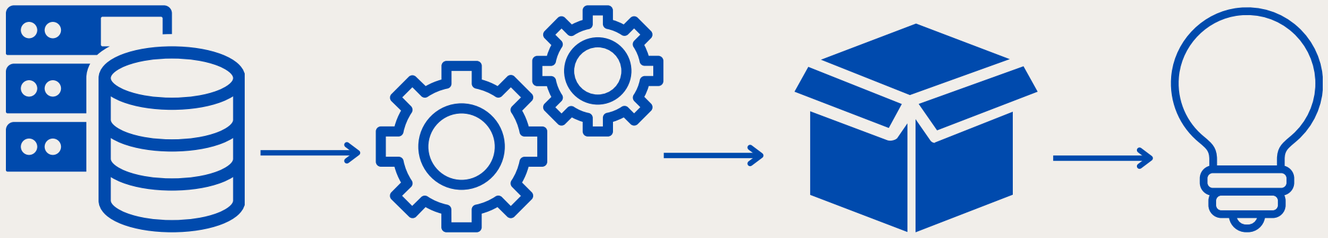
- AUC = 0.63 → Train y test similares
- Permite usar validación cruzada estándar sin riesgo de sobreajuste



A partir de 44 variables → **156 features**

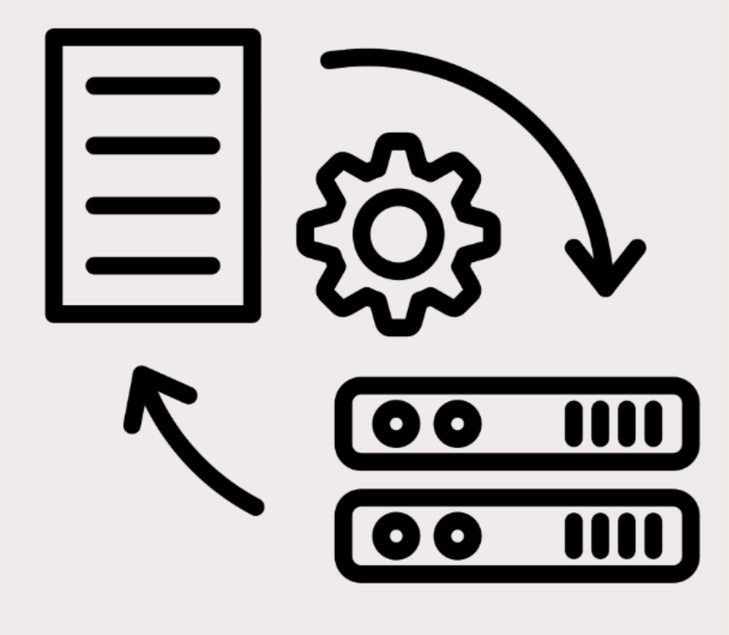
Tipos de features:

- **Numericas**: log(precio), ratios (precio actual/base), diferencias, unidades vendidas
- **Temporales**: duración del listing, mes/día de publicación/actualización
- **Textuales**: longitud del título, proporción de mayúsculas/dígitos, palabras clave (‘nuevo’, ‘usado’)
- **Vendedor y atributos**: reputación, frecuencia de publicaciones, tienda oficial, envío, garantía, cantidad de imágenes
- **Interacciones**: combinaciones de features que capturan relaciones no lineales



# Feature Engineering

# Procesamiento de texto y categóricas



- **TF-IDF**: transforma títulos en vectores numéricos
  Detecta palabras clave y patrones semánticos

- **Label Encoding**: convierte variables categóricas en números
  Permite al modelo “entender” el contenido textual sin perder información

# Entrenamiento y validación

Validación cruzada estratificada 10 folds → mantiene proporción de clases

Modelos entrenados:

- **LightGBM**: rápido y eficiente
- **XGBoost**: alta capacidad de generalización
- **CatBoost**: maneja categóricas eficientemente

Parámetros: learning rate 0.02, profundidad 9, early stopping

5 semillas → promedio de resultados → **AUC ≈ 0.9985**

**ENSEMBLE PONDERADO**

Combinación de modelos:

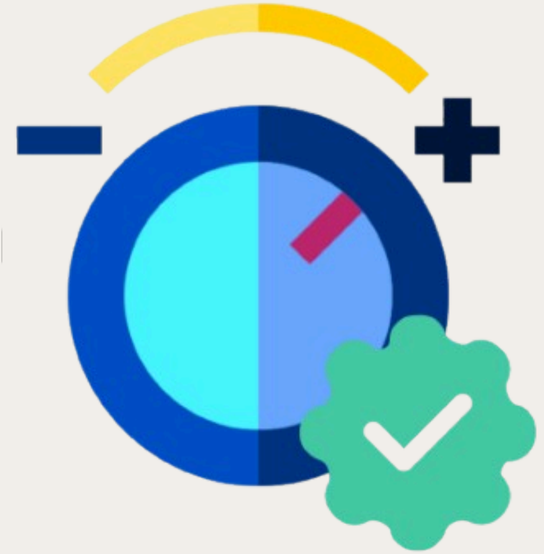
- LightGBM: 35%
- XGBoost: 35%
- CatBoost: 30%

suaviza errores individuales → predicción más robusta

# Mejoras adicionales

1. **Pseudo-Labeling**
   - Agrega predicciones confiables del test como etiquetas
   - Dataset ampliado +23 filas → mejor generalización


1. **Optimización de threshold**
   - Threshold óptimo: 0.4836
   - Mejora F1-score → Accuracy ~98%



# Resultados

**Modelo final**: Ensemble + pseudo-labeling + threshold optimizado

**AUC validación**: 0.9985

**Competencia Kaggle**:

- Score público: 0.90700
- Score privado: 0.90911

**Posibles mejoras futuras:**

- Búsqueda de hiperparámetros más fina
- Stacking con meta-modelo
- Análisis más profundo de importancia de variables (SHAP Values)

# Muchas gracias