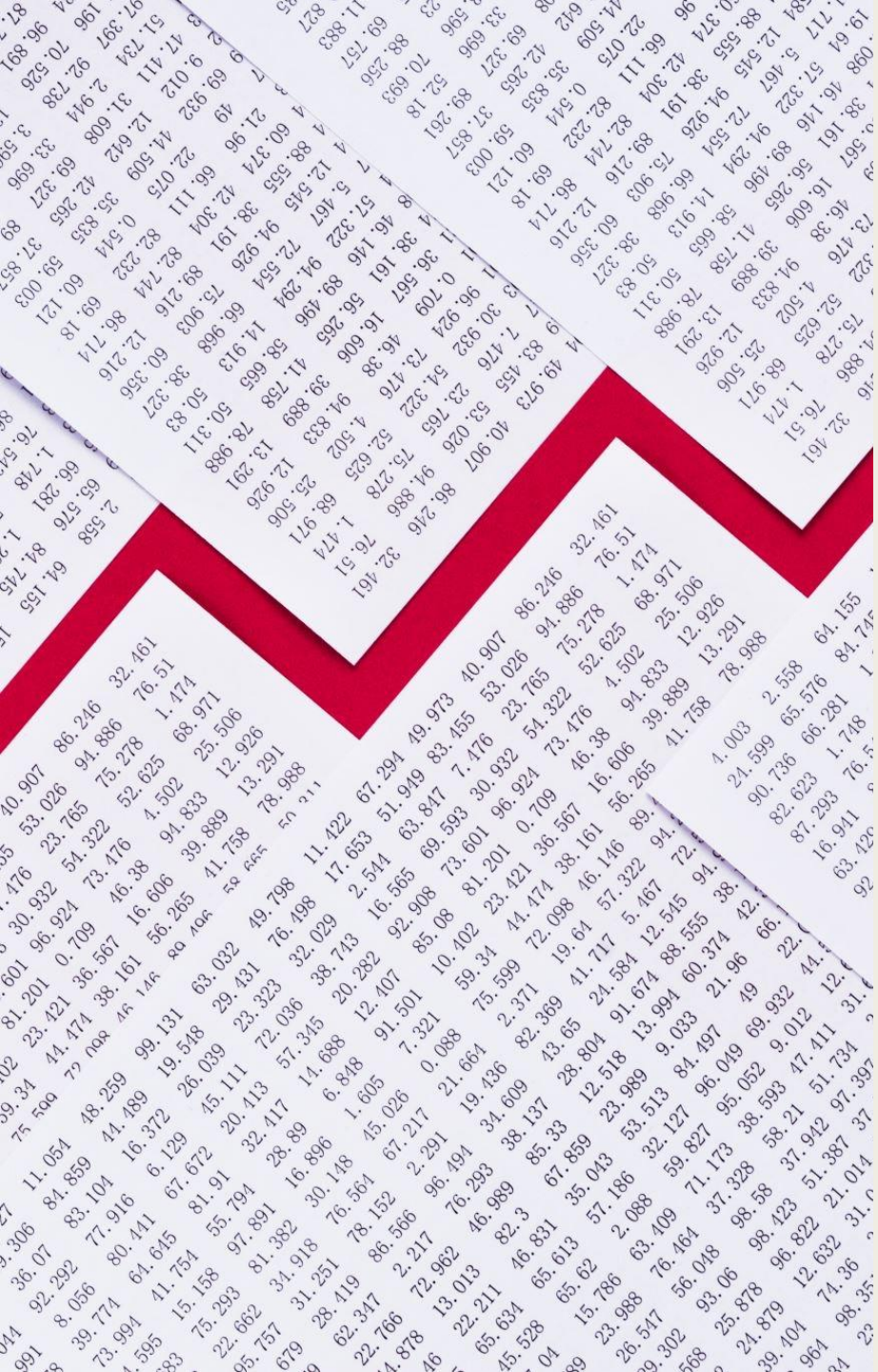
PREDICCIÓN DE REVENUE EN VENTAS DE AMAZON MEDIANTE REGRESIONES Y PENALIZACIONES

LOUIS LEONE BLANCAS MOREYDA

SANTIAGO ESCUTIA RIOS

EMILIO NAVARRO CABRALES



OBJETIVOS

- General:

Determinar el potencial de ingresos de las transacciones en Amazon mediante modelos de aprendizaje supervisado, para identificar los factores comerciales que garantizan la rentabilidad.

- Específicos:

Procesar y transformar datos crudos (fechas y categorías) en variables numéricas aptas para modelos de regresión.

Evaluar y comparar el rendimiento de modelos con regularización (**Lasso, Ridge, ElasticNet**) para asegurar la capacidad de generalización del modelo.

Analizar la significancia de los coeficientes para cuantificar el impacto real de los descuentos y el volumen de ventas en el margen final.

MARCO TEÓRICO

Tema Elegido: Análisis predictivo del revenue en e-commerce, enfocado en el uso de modelos lineales avanzados para comprender la dinámica financiera de transacciones en Amazon.

Regresión Lineal y Polinomial: La regresión lineal estima relaciones directas entre variables, mientras que la regresión polinomial permite capturar relaciones no lineales mediante la inclusión de términos de orden superior.

Interacción de Factores: Permite evaluar cómo el efecto de una variable (por ejemplo, descuento) varía en función de otra (como categoría), capturando dependencias estructurales entre predictores.

Significancia de Factores: Se evalúa mediante el p-value para determinar si los coeficientes estimados tienen impacto estadísticamente significativo sobre el revenue.

Regularización (Ridge, Lasso, ElasticNet): Métodos que incorporan penalizaciones a la función de pérdida para reducir el sobreajuste. Lasso favorece la selección de variables, Ridge mejora la estabilidad ante multicolinealidad y ElasticNet combina ambos enfoques.

ANÁLISIS DEL DATASET

Origen: Dataset público de ventas de Amazon	**Muestra:** 50,000 transacciones	**Variables:** Precio, Cantidad, Descuento, Rating, Categoría y Fecha	**Target:** Revenue
Preparación →	Fecha, Año y Mes (estacionalidad)	Eliminación de IDs	0% valores nulos

TRANSFORMACIONES NECESARIAS

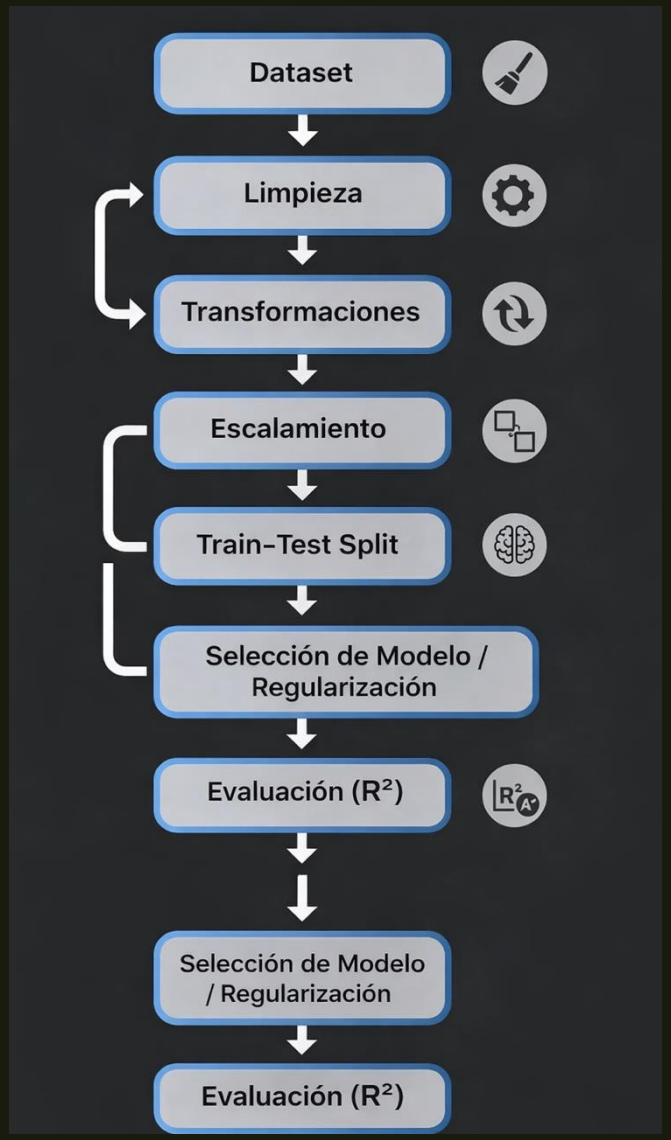
Ingeniería de Características (Feature Engineering)

Conversión Temporal: Transformación de la variable order_date a formato *datetime* para extraer **Año** y **Mes**. Esto permite capturar la estacionalidad de las ventas.

Selección de Atributos: Eliminación de identificadores únicos (Order ID, Product ID) que no aportan poder predictivo y generan ruido algorítmico.

Auditoría de Calidad: Verificación y limpieza de datos, resultando en un dataset con **0% de valores nulos** para los 50,000 registros.

PIPELINE


Dataset
Limpieza
Transformaciones
Escalamiento
Train-Test Split
Selección de Modelo / Regularización
Evaluación (R^2)
Selección de Modelo / Regularización
Evaluación (R^2)

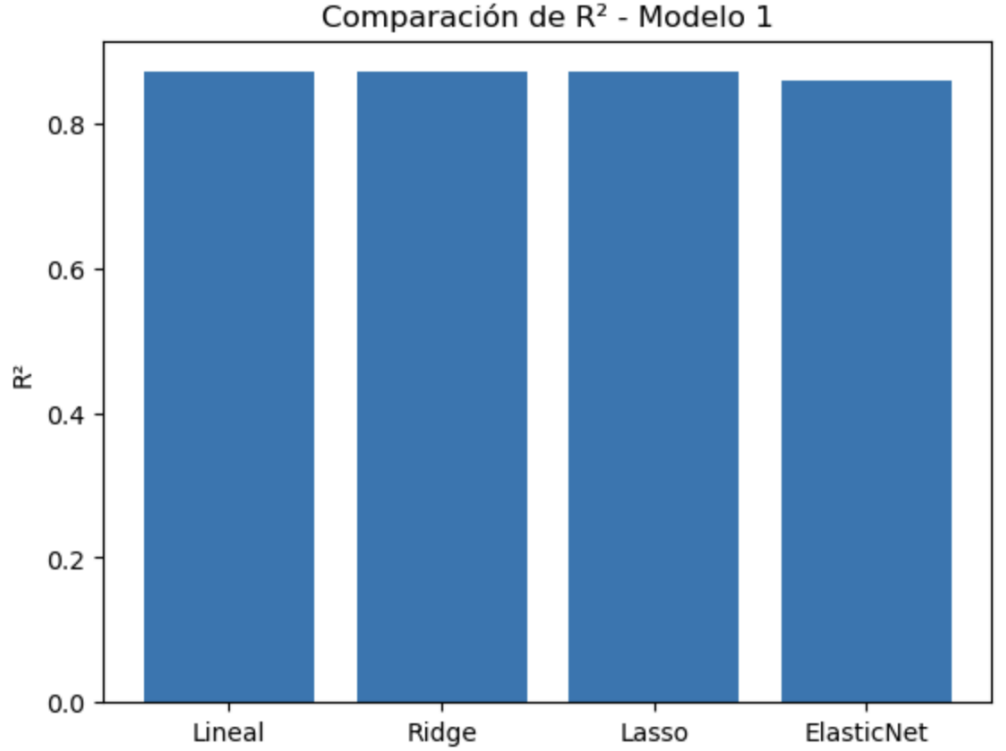
ESTRATEGIA DE MODELADO

Enfoque: Se definieron 3 modelos base, evaluando 4 versiones para cada uno:

- **Sin penalización** (Lineal simple).
- **Ridge** (Estabilidad).
- **Lasso** (Simplicidad).
- **ElasticNet** (Híbrido).

Validación: Split 80% entrenamiento / 20% prueba.

Comparativa de Desempeño: ¿Qué modelo predice mejor?



Modelo	R^2 Score (Test)	Estatus
Regresión Lineal	0.87069	Base
Lasso (L1)	0.87075	GANADOR
Ridge (L2)	0.87069	Estable
ElasticNet	0.85890	Menor ajuste

ANÁLISIS DE SIGNIFICANCIA ¿Qué factores impulsan el Revenue?

Volumen de Ventas (Coef: +220.7): Es el predictor dominante. Por cada unidad extra vendida, el impacto en el ingreso es masivo y positivo.

Precio Unitario (Coef: +2.6): Tiene una relación positiva directa, aunque su peso es menor comparado con la cantidad.

Impacto del Descuento (Coef: -7.6): El porcentaje de descuento afecta negativamente el ingreso por unidad, indicando que las rebajas deben ser estratégicas para no erosionar el margen.

P-Value: Todos los factores presentan un valor **< 0.05**, confirmando su relevancia estadística.

CONCLUSIONES
Principales Hallazgos

Alta Precisión: El modelo explica el **87% de la variabilidad** de los ingresos (R2=0.8707).

Modelo Ganador: Lasso (L1) resultó ser la técnica más eficiente por su equilibrio entre precisión y simplicidad.

Motor de Ventas: El éxito financiero depende principalmente del **volumen vendido** (quantity_sold), siendo el factor más influyente.

Riesgo de Descuento: Los descuentos tienen un **impacto negativo** directo; solo son rentables si disparan masivamente el volumen de ventas.

Repositorio del Proyecto

El código completo, notebook y dataset se encuentran disponibles en:

GitHub:

https://github.com/sanesc21/Repositorio-del-Proyecto--P01---Regresion/tree/main

Repositorio público con:

- Notebook en Jupyter
- Dataset

Referencias Bibliográficas

- **Dataset:** Hussain, A. (2024). *Amazon Sales Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/aliiihussain/amazon-sales-dataset
- **Librería de Modelado:** Scikit-learn: Machine Learning in Python. Recuperado de https://scikit-learn.org/ (Utilizado para Lasso, Ridge y ElasticNet).
- **Procesamiento de Datos:** Pandas: Data structures for Python. Recuperado de https://pandas.pydata.org/ (Utilizado para limpieza y transformación).