INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

Tecnicatura Superior en
Ciencia de Datos e Inteligencia Artificial

DealData
Análisis Predictivo de Calificaciones de Productos en Amazon Mediante Regresión Logística

Proyecto ABP
Materias: Ciencia de Datos II y Estadísticas y Exploración de Datos II

Integrantes:

- Carlos Direni
- Nicolás Allende
- Melania Ligorria
- Guadalupe Mendoza
- Miguel Rojas Medina
- Emanuel Guaraz
- Lucas Ryser
- Juan Clavijo

Docentes Guía:

- Nahuel Pratta
- Marcos Ugarte

AÑO: 2025

Dirección General de EDUCACIÓN TÉCNICA Y FORMACIÓN PROFESIONAL
Ministerio de EDUCACIÓN
CÓRDOBA Seguimos haciendo

Proyecto ABP

Integrantes del equipo

Nombre de los integrantes	Usuario GitHub
Melania Ligorria	https://github.com/mel-ligorria
Miguel Rojas	https://github.com/Migueerm
Guadalupe Mendoza	https://github.com/Guadamendoza
Carlos Direni	https://github.com/Cdireni1
Nicolás Allende Olmedo	https://github.com/AllendeNicolas
Emanuel Guaraz	https://github.com/JEmanuelG
Lucas Ryser	https://github.com/lucasryser6
Juan Clavijo	https://github.com/juancla001

Link repositorio GitHub

https://github.com/mel-ligorria/DealData

Acerca del Data Set

Este conjunto de datos contiene información sobre más de 1000 calificaciones y reseñas de productos de Amazon, según detalles que aparecen en el sitio web oficial de la empresa.

El Data Set, posee un considerable número de registros, y una amplia cantidad de variables de interés, que podemos utilizar para realizar distintos tipos de análisis sobre los productos más solicitados de esta tienda reconocida a nivel mundial.

Resumen del Proyecto

- **Tipo:** Tecnológico y de Investigación.
- **Ejes Conceptuales:** el proyecto abarca el ciclo completo de un análisis de datos, incluyendo: preprocesamiento de datos (Data Cleaning), Análisis Exploratorio de Datos (EDA), ingeniería de variables, modelado predictivo con regresión logística y evaluación de modelos de clasificación.

Problemática y Fundamentación

- **Problemática:** en el competitivo mercado del comercio electrónico, las calificaciones de los productos son un indicador crucial de la satisfacción del cliente y el éxito comercial. Sin embargo, no siempre es evidente qué factores (precio, descuentos, popularidad) influyen más en que un producto sea bien valorado. Este proyecto aborda la necesidad de identificar y cuantificar el impacto de estas características sobre la probabilidad de que un producto reciba una alta calificación en Amazon.
- **Fundamentación e hipótesis:** la capacidad de anticipar la recepción de un producto permite a vendedores y empresas optimizar sus estrategias. Este proyecto es fundamental para el perfil del Técnico en Ciencia de Datos, ya que aplica un ciclo completo de análisis, desde la manipulación de datos brutos hasta la generación de insights accionables.
    - **Impacto de negocio:** la capacidad de predecir la calificación permite optimizar la **estrategia de precios**. Si un precio más alto se asocia a mejores ratings, se puede evitar una estrategia de descuentos agresivos que podría perjudicar la percepción de calidad. Estos insights son claves para la **gestión de inventario y marketing**, permitiendo enfocar los recursos en productos con alta probabilidad de éxito.

- **Hipótesis Nula (H_0):** el precio, el descuento y la cantidad de valoraciones no

  tienen un efecto estadísticamente significativo en la probabilidad de que un producto sea “altamente calificado”.
- **Hipótesis Alternativa (H_1):** se espera que un mayor número de valoraciones y un precio más alto tengan un **efecto positivo** en la probabilidad de una alta calificación, mientras que un mayor porcentaje de descuento podría tener un **efecto negativo**

Objetivos del proyecto

Objetivo general

Desarrollar y evaluar un modelo de regresión logística para predecir si un producto tiene una alta probabilidad de ser calificado con más de 4.0 estrellas, basándose en su precio, porcentaje de descuento y número de valoraciones.

Objetivos específicos

1. Realizar un preprocesamiento exhaustivo del conjunto de datos.
2. Ingeniar una variable objetiva binaria ("is_highly_rated") para la clasificación.
3. Seleccionar y transformar las variables predictoras más relevantes.
4. Construir y validar un modelo de regresión logística.
5. Evaluar el rendimiento del modelo mediante métricas clave (Accuracy, Matriz de Confusión, Precisión, Recall y AUC).
6. Interpretar los coeficientes del modelo para extraer conclusiones de negocio.

Metodología y fases del proyecto

Para la ejecución del proyecto, se siguió un flujo de trabajo estructurado, justificando cada decisión técnica para garantizar la robustez y reproducibilidad del análisis. Se utilizó una división estándar de **80/20 para los datos de**

entrenamiento y prueba, proporción que permite entrenar el modelo con una cantidad sólida de datos (80%) mientras se reserva un subconjunto significativo e independiente (20%) para evaluar su capacidad de generalización.

El proyecto se desarrolló en las siguientes seis fases:

1. **Preparación del entorno y datos:** se utilizó un conjunto de datos de Kaggle ("Amazon Product Reviews Dataset"). El entorno de trabajo se configuró en un Jupyter Notebook, importando las librerías necesarias.
2. **Análisis Exploratorio de Datos (EDA):** se realizó un análisis inicial para entender la estructura de los datos, identificar tipos de datos, valores faltantes y la distribución de las variables numéricas.
3. **Curación y limpieza de datos:** se aplicó un riguroso proceso de limpieza
    a. Se eliminaron todas las filas con valores nulos (dropna).
    b. Se convirtieron las columnas de texto con valores numéricos (discounted_price, actual_price, etc.) a formato numérico, eliminando símbolos.
    c. Se identificaron y eliminaron valores atípicos (outliers) utilizando el método del Rango Intercuartílico (IQR), lo que redujo el dataset de 1463 a 999 registros para el modelado.
4. **Ingeniería de variables:**
    a. Se creó la variable objetivo binaria is_highly_rated (1 si el rating era > 4.0), necesaria para el modelo de clasificación.
    b. Se aplicó una transformación logarítmica (log1p) a las variables con sesgo positivo (actual_price, rating_count) para normalizar su distribución.
5. **Modelado y evaluación:**
    a. Se seleccionaron las variables predictoras más relevantes (log_actual_price, discount_percentage, rating_count_log) mediante métodos de selección Forward y Backward.
    b. Se verificó la ausencia de multicolinealidad usando el Factor de

Inflación de la Varianza (VIF).

c. Se entrenó un modelo de regresión logística con statsmodels y se validó con sklearn(solucionador L-BFGS).

d. La evaluación del modelo arrojó un **Accuracy del 64.85%** y un **AUC de 0.66**, indicando un rendimiento predictivo moderado pero superior al azar.

6. **Interpretación y conclusiones:** se analizaron los coeficientes y p-values del modelo para extraer conclusiones significativas sobre el impacto de cada variable.

Cronograma de actividades

Objetivo Específico	Acciones, Recursos y Tiempos
1. Preprocesamiento del Dataset	**Acciones:** carga, limpieza de nulos/duplicados, depuración de columnas numéricas y tratamiento de outliers. **Recursos:** Python, Jupyter, Pandas, NumPy. **Tiempo:** semana 1.
2. Creación de variable objetivo	**Acciones:** definir umbral, crear la columna is_highly_rated. **Recursos:** Pandas. **Tiempo:** semana 1.

3. Selección y transformación de predictores	**Acciones:** análisis de correlación y VIF, transformación logarítmica, selección con métodos Forward/Backward. **Recursos:** Pandas, Statsmodels, Matplotlib. **Tiempo:** semana 2.
4. Construcción y entrenamiento del modelo	**Acciones:** división de datos (entrenamiento/prueba), ajuste de modelos statsmodels y sklearn. **Recursos:** Statsmodels, Scikit-learn. **Tiempo:** semana 2-3.
5. Evaluación del modelo	**Acciones:** Cálculo de Accuracy, Precisión, Recall, F1-Score, Matriz de Confusión y AUC. **Recursos:** Scikit-learn, Matplotlib. **Tiempo:** semana 3.
6. Interpretación y conclusiones	**Acciones:** análisis de resultados estadísticos, interpretación de coeficientes y elaboración de conclusiones finales. **Recursos:** herramientas de redacción. **Tiempo:** semana 4.

Herramientas utilizadas

Para la realización de este proyecto se emplearon las siguientes tecnologías y librerías del ecosistema de Python:

- **Lenguaje:** Python
- **Entorno:** Jupyter Notebook (Google Colab)
- **Manipulación de Datos:** Pandas, NumPy
- **Visualización:** Matplotlib, Seaborn
- **Modelado Estadístico:** Statsmodels, Scikit-learn
- **Control de Versiones:** GitHub

Conclusiones finales

El modelo de regresión logística desarrollado es estadísticamente significativo y confirma que las variables seleccionadas influyen de manera predecible en la calificación de los productos en Amazon.

Conclusiones principales

- Se confirma que la **cantidad de valoraciones es el predictor con el impacto positivo más fuerte**, indicando que la popularidad y la prueba social de un producto son factores clave para su éxito y la probabilidad de recibir una alta calificación.
- Se descubre una **relación negativa pero significativa con el porcentaje de descuento**. Esto sugiere que, si bien los descuentos pueden impulsar las ventas, las rebajas agresivas podrían estar asociadas a una menor percepción de calidad por parte de los consumidores, afectando negativamente su calificación.
- El **precio del producto** también muestra una relación positiva, lo que indica que, controlando los otros factores, un precio más alto se asocia ligeramente a una mayor probabilidad de ser bien calificado.

Entregables del proyecto

- Un modelo de regresión logística entrenado y funcional.
- Un cuaderno de Jupyter con el código documentado.
- El presente informe ejecutivo.

Limitaciones y trabajo futuro

Se reconoce que, si bien el proyecto alcanzó sus objetivos, existen limitaciones inherentes y oportunidades claras para futuras mejoras.

1. **Limitaciones**
    a. El **poder predictivo del modelo es moderado (AUC de 0.66)**, lo que sugiere que las variables utilizadas, aunque relevantes, no capturan toda la complejidad que define la satisfacción del cliente.
    b. El análisis **se limita a variables numéricas** y no considera el valioso contenido textual de las reseñas de los usuarios, donde reside una gran cantidad de información cualitativa.
2. **Trabajo Futuro** para superar estas limitaciones y construir un modelo más preciso, se proponen las siguientes líneas de trabajo:
    a. **Análisis de Sentimiento:** Aplicar técnicas de **Procesamiento de Lenguaje Natural (NLP)** a las reseñas de los productos para extraer sentimientos (positivos, negativos, neutros) y temas recurrentes, utilizándolos como nuevas y potentes variables predictivas.
    b. **Enriquecimiento del Dataset:** Incorporar **información adicional**, como la categoría del producto, la marca o la reputación del vendedor, para añadir más contexto y mejorar la capacidad explicativa del modelo.
    c. **Exploración de otros algoritmos:** Implementar y evaluar otros modelos de clasificación (ej. Random Forest, Gradient Boosting) para comparar su rendimiento con la regresión logística y determinar si pueden capturar relaciones no lineales en los datos de manera más efectiva.

Producto final y conclusiones

El producto final es un modelo de regresión logística validado y documentado que, aunque presenta un poder predictivo moderado (Accuracy ≈ 65%, AUC ≈ 0.66), es estadísticamente robusto y superior a una clasificación aleatoria.

Conclusiones principales:

- Se confirma que la **cantidad de valoraciones** es el predictor más fuerte de una calificación alta, seguido por el precio del producto.
- Se descubre una **relación negativa entre el porcentaje de descuento y la calificación**, sugiriendo que los descuentos agresivos podrían asociarse a una menor percepción de calidad.

Entregables del proyecto:

1. Un **modelo de regresión logística entrenado** y serializado, capaz de realizar predicciones.
2. Un **cuaderno de Jupyter (Notebook)** con todo el código del análisis documentado.
3. El **presente informe ejecutivo resume** el problema, la metodología y los hallazgos.

Segundo cuatrimestre

Dealdata, es el resultado de la unificación de tres grupos diferentes, respecto de los equipos formados en el primer cuatrimestre. Por lo cual decidimos trabajar y afianzarnos como equipo, utilizando una Base de Datos diferente, ya que nuestros trabajos anteriores eran muy diferentes, y por cuestiones de incompatibilidad, no podíamos continuar con alguno de los Data Set anteriores. Por esto decidimos trabajar con un Data Set de Ventas de Amazon, el cual obtuvimos de la página de Kaggel (https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset), el cual contaba con los requisitos específicos de poder visualizar en el mismo, variables

preferentemente cuantitativas para de esta manera poder realizar los análisis estadísticos pertinentes. Nuestro primer paso fue la normalización del data set, eliminando datos erróneos o contradictorios, y reduciendo los Outliers, de la mejor manera posible.

En el Colab de Google, se realizó un análisis de datos, aplicando diferentes métodos y técnicas centradas en comprender ciertos aspectos de los productos de Amazon.

Primero, se exploró la relación entre la cantidad de calificaciones recibidas por un producto y su calificación promedio. Se utilizaron diferentes métodos de correlación para ver si los productos más populares tendían a tener mejores calificaciones.
Luego, se aplicó una técnica llamada ANOVA para comparar las calificaciones promedio de los productos entre diferentes categorías. Esto ayudó a identificar si el tipo de producto influye en qué tan bien es calificado por los usuarios.

Finalmente, se desarrolló un modelo de regresión logística multivariable. El objetivo de este modelo fue predecir la probabilidad de que un producto sea considerado “altamente calificado”, basándose en características como su precio, el descuento ofrecido y la cantidad de calificaciones. Se evaluó qué tan bien el modelo podía hacer estas predicciones.

Puede analizar, visualizar el desarrollo y las conclusiones obtenidas del trabajo, en el Repositorio del Grupo, y en la carpeta ABP del mismo.