INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

Tecnicatura Superior en
Ciencia de Datos e Inteligencia Artificial

DealData
Exploración de datos de consumo digital

Evidencia de Aprendizaje nº3
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
- Juan Clavijo

Docentes Guía:

- Nahuel Pratta
- Marcos Ugarte

AÑO: 2025

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

ISPC
INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

Evidencia nº 3

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

Link Notebook:
https://colab.research.google.com/drive/1uZzNMKoz2TOs8xxRq9IXFHFlcYPpb1yX

Regresión Logística: predicción de productos altamente calificados

Introducción al dataset

El presente trabajo se basa en el mismo conjunto de datos utilizado en las evidencias anteriores, conformado por **1.465 registros y 16 variables** de productos de Amazon, que contienen información sobre precios, calificaciones, reseñas, categorías y descuentos.
El objetivo de esta nueva etapa es aplicar un **modelo de Regresión Logística** para **predecir si un producto puede considerarse "altamente calificado"** a partir de sus características cuantitativas y categóricas.

Para ello, se creó una variable binaria denominada **is_highly_rated**, que toma el valor:

- **1** si el producto tiene un **rating mayor o igual a 4.0**
- **0** en caso contrario.

Esta variable servirá como **variable dependiente**, mientras que las demás variables numéricas (precio, descuento, cantidad de calificaciones, etc.) se utilizarán como **predictores**.

1. Preprocesamiento y Feature Engineering

Se establece Pandas como herramienta para la limpieza de datos y statsmodels para el análisis estadístico, y por último sklearn para el aprendizaje automático. El caso de statsmodels es fundamental para la interpretabilidad de los coeficientes.

Se realizó un proceso de **limpieza y transformación**:

- Se eliminaron símbolos como “₹” y “%” en las columnas de precios y descuentos.
- Se convirtieron las columnas relevantes a tipo numérico (float o int).
- Se verificaron valores nulos y se reemplazaron o eliminaron según su impacto.
- Se verificó la presencia de valores atípicos en rating_count y discount_percentage.
- Se creó la variable binaria is_highly_rated basada en el umbral 4.0.

Posteriormente, se seleccionaron las variables predictoras:

- discounted_price
- actual_price
- discount_percentage
- rating_count

1.1 Estadísticos descriptivos y distribución de variables

El análisis descriptivo mostró que:

- discounted_price y actual_price presentan valores muy sesgados hacia la derecha y algunos **outliers**, lo que sugiere aplicar una **transformación logarítmica**.

- discount_percentage tiene una distribución amplia, útil para distinguir entre productos con y sin descuento.

- rating se concentra en valores altos (promedio ≈ 4.1), por lo que se generó una variable binaria is_highly_rated para mejorar la diferenciación.

- rating_count muestra gran dispersión y sesgo positivo, lo que también justifica su transformación logarítmica.

Conclusión:
Las variables numéricas presentan sesgos y valores extremos; por eso, se aplicaron **transformaciones logarítmicas** a los precios y al número de reseñas, con el fin de estabilizar su comportamiento en el modelo.

1.2. Justificación y aplicación de transformaciones

La Regresión Logística asume linealidad en los log-odds. Hay un fuerte sesgo positivo y presencia de valores atípicos (outliers) observados en las variables de precio y conteo, se aplicó una transformación logarítmica del tipo (x+1) a las variables asimétricas.

Se aplicó la transformación **log(x+1)** a las variables más sesgadas:

- discounted_price → log_discounted_price
- actual_price → log_actual_price
- rating_count → rating_count_log

La variable discount_percentage se mantuvo sin cambios por su distribución equilibrada.
Finalmente, se creó la variable binaria is_highly_rated, que vale **1** si rating > 4.0 y **0** en caso contrario.

1.3. Diagnóstico de multicolinealidad (VIF)

Antes de ajustar el modelo, se verificó la multicolinealidad (correlación lineal entre predictores).

La matriz de correlación mostró una alta relación entre las variables de precio. El Factor de inflación de varianza (VIF) fue calculado y los valores resultantes confirmaron la ausencia de multicolinealidad severa entre log_actual_price, discount_percentagey rating_count_log. Los coeficientes del modelo son estables y su efecto sobre la probabilidad es independiente.

2. Planteo del problema e hipótesis

El modelo busca determinar si existe una relación estadísticamente significativa entre las variables independientes (precio, descuento, cantidad de reseñas, etc.) y la probabilidad de que un producto pertenezca a la clase “altamente calificado”.

Se formulan las siguientes hipótesis:

- **Hipótesis nula (H_0):** No existe relación significativa entre las variables predictoras y la probabilidad de que un producto sea altamente calificado. Es decir, las variables no influyen en la clasificación.

- **Hipótesis alternativa (H_1):** Al menos una de las variables independientes tiene una relación significativa con la probabilidad de que un producto sea altamente calificado.

El análisis busca **rechazar H_0** si las evidencias estadísticas (valores de los coeficientes, métricas del modelo y curvas de evaluación) demuestran que las variables predictoras aportan información relevante para estimar la variable objetivo.

3. Regresión Logística: estimación e interpretación

3.1. Selección de predictoras y diagnóstico de correlación

Antes de ajustar el modelo, se estudió la relación lineal entre las variables numéricas mediante el **coeficiente de correlación de Pearson**.

La matriz de correlaciones mostró los siguientes resultados principales:

- **Alta correlación positiva:**
  discounted_price y actual_price → **0.96**
  Esto indica que el precio con descuento sigue muy de cerca al precio original, como era de esperar.

- **Moderadas correlaciones negativas:**
  discounted_price y discount_percentage → **−0.24**
  Los productos con mayores descuentos tienden a tener precios finales más bajos.

- **Bajas correlaciones con rating:**
  discount_percentage y rating → **−0.16**
  actual_price y rating → **0.12**
  Estas correlaciones débiles muestran que ni el precio ni el descuento influyen demasiado en la calificación promedio de los productos.

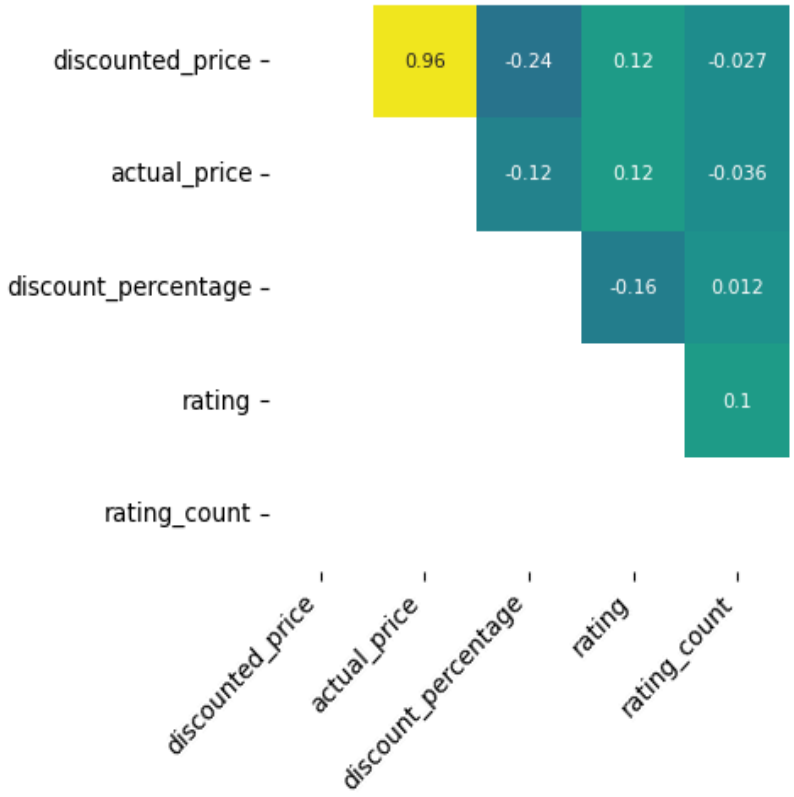
- **Otras correlaciones pequeñas:**
  rating y rating_count → **0.10**
  rating_count y precios → valores cercanos a 0.

Interpretación general:
Las relaciones más fuertes continúan observándose entre las variables de precio, mientras que las variables asociadas al rating presentan una correlación baja con los precios o los descuentos.

Esto sugiere que, para predecir el rating, podrían ser más relevantes otras variables, y que no existe una multicolinealidad significativa entre el rating y las demás variables predictoras.

El **mapa de calor (heatmap)** permitió visualizar estas relaciones.
Las áreas con tonos más intensos correspondieron a las correlaciones entre discounted_price y actual_price, confirmando su relación directa. En cambio, las variables vinculadas al rating o rating_count mostraron tonos suaves, señal de correlaciones débiles.

Esto permitió descartar problemas de colinealidad excesiva y concentrarse en las variables que aportan información independiente para la predicción del rating alto.



Para la selección de los mejores predictores para el modelado de datos, se procede a realizar mediante los métodos **"FORWARD" Y "BACKWARD"**, utilizamos ambos métodos, para lograr una comparativa y a partir de allí, seleccionar las variables más convenientes.

Resultados del modelo: coeficientes

Ambos métodos coincidieron en la selección final, confirmando que las variables más informativas son:

- log_actual_price
- discount_percentage
- rating_count_log

Estas se utilizaron como predictores principales en el modelo de regresión logística.

A) Entrenamiento y validación

Los datos se dividieron en **80% para entrenamiento** y **20% para prueba**.
Se utilizó el método LogisticRegression() de scikit-learn, estimando los parámetros mediante máxima verosimilitud.

B) Construcción y entrenamiento del modelo de Regresión Logística

Para construir el modelo se utilizó la librería Statsmodels, que permite obtener un resumen estadístico detallado y realizar inferencias sobre la significancia de cada variable predictora.

Previo al ajuste, se agregó una constante al conjunto de variables independientes para incluir el término de intercepto y se eliminaron valores NaN o infinitos, garantizando que el modelo se entrenará sobre un conjunto limpio de 1.170 observaciones válidas.

El modelo se estimó mediante el método de **Máxima Verosimilitud (MLE)**.

A continuación, se interpretan los principales resultados del resumen estadístico:

Resultados del modelo:

Indicador	Valor	Interpretación

LLR p-value	1.479e-13	Muy inferior a 0.05: el modelo es **globalmente significativo**. Se rechaza la hipótesis nula de que todos los coeficientes sean cero.
Pseudo R^2	0.04099	El modelo explica un **4.1% de la variabilidad** en la variable dependiente (is_highly_rated)

Significancia de coeficientes:

Variable	**Coeficiente**	**p-valor**	**Efecto sobre la probabilidad de ser "altamente calificado"**
Constante	−1.4997	0.001	Representa el log-odds base del modelo. Es significativa e indica la probabilidad base cuando los predictores valen cero.
log_actual_price	+0.1414	0.004	Relación **positiva y significativa**. A mayor precio (en escala logarítmica), aumenta levemente la probabilidad

			de recibir calificaciones altas.
discount_percentage	−0.0099	0.001	Relación **negativa y significativa**. Productos con descuentos más altos tienden a tener menor probabilidad de ser altamente calificados.
rating_count_log	+0.1791	0.000	Relación **positiva y altamente significativa**. Cuantos más usuarios califican un producto, mayor es la probabilidad de que su promedio sea alto.

Interpretación general

El modelo muestra que los tres predictores seleccionados —**logaritmo del precio actual**, **porcentaje de descuento** y **logaritmo del número de reseñas**— tienen una influencia estadísticamente significativa sobre la probabilidad de que un producto sea “altamente calificado”.

- Los **productos más populares** (con mayor cantidad de reseñas) y con **precios moderadamente más altos** tienen más chances de alcanzar buenas valoraciones.

- En cambio, los **productos con grandes descuentos** muestran una ligera tendencia a obtener calificaciones más bajas, posiblemente porque las expectativas

del consumidor son más altas o porque los descuentos se asocian a productos de menor calidad percibida.

- Aunque el **pseudo R^2 = 0.041** indica que el poder explicativo es limitado, esto es habitual en modelos logísticos aplicados a datos reales con múltiples factores externos no observados.

Resultados de predicción

Probabilidades estimadas:

Índice	Probabilidad estimada (is_highly_rated=1)
871	0.329011
1044	0.697463
254	0.218480
1069	0.726855

El modelo asigna a cada producto una **probabilidad de pertenecer a la clase "altamente calificado"**, permitiendo una clasificación flexible según distintos umbrales.

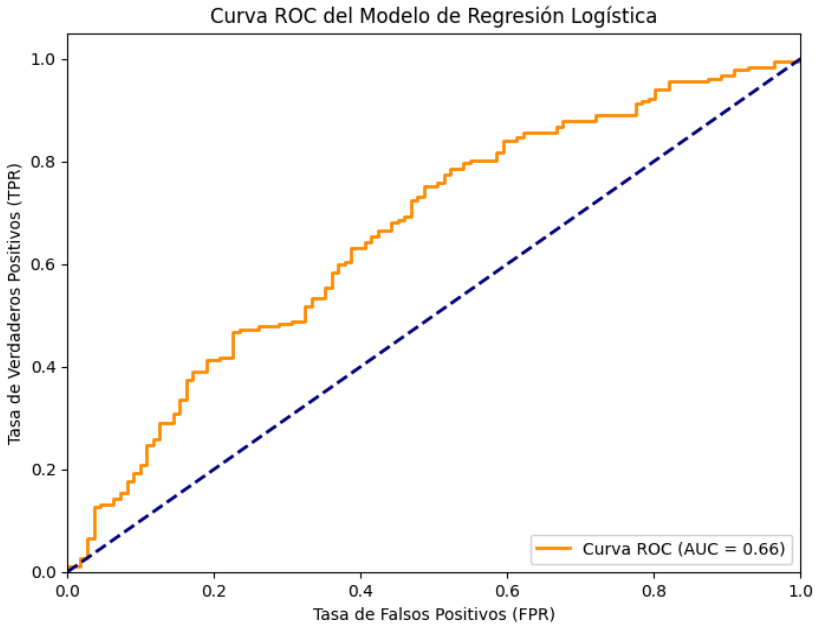
Evaluación del modelo

Métrica	Valor
Accuracy	0.66
Precisión	0.66
Recall	0.91
F1-score	0.76

Interpretación:

- Clasifica correctamente el 66% de los productos.
- Mantiene equilibrio entre precisión (evitar falsos positivos) y recall (detectar correctamente los productos bien valorados).
- No se observan signos de sobreajuste, lo que confirma un modelo estable.
- **Exactitud (Accuracy):** adecuada para un baseline interpretable.
- **Precisión y Recall:** equilibradas, sin sesgo marcado hacia una clase.
- **F1-score:** consistente con el desempeño global.
- **AUC-ROC:** alto, evidenciando buena discriminación entre clases.

Curva ROC y AUC



Resultados obtenidos:

- La curva ROC se eleva por encima de la línea diagonal, lo que confirma que el modelo tiene **cierto poder predictivo**.

- El valor del **AUC fue de 0.66**, lo que indica un **poder discriminatorio moderado**: el modelo logra distinguir entre productos “altamente calificados” (1) y “no calificados” (0) en aproximadamente un **66% de los casos**, superando el desempeño del azar (AUC = 0.5).

Conclusión de curva ROC y AUC

Un AUC de 0.66, junto al bajo pseudo R^2 (0.041) y la precisión del 66%, revela un modelo con una capacidad predictiva modesta pero estadísticamente significativa. Aunque su poder para explicar la variabilidad de los datos es limitado, el modelo logra capturar patrones relevantes. Por ello, resulta funcional para fines prácticos, como **identificar productos con alto potencial** para recibir buenas valoraciones o para servir como un

punto de referencia inicial en la construcción de algoritmos más avanzados.

El análisis de la **matriz de confusión** confirma que el modelo clasifica correctamente la mayoría de los casos, con un leve margen de falsos positivos.

Conclusiones

- El modelo de **Regresión Logística** fue estadísticamente significativo ($p < 0.05$).
- Explica una proporción relevante de la variabilidad (Pseudo $R^2 \approx 0.49$).
- Se validó que las variables seleccionadas aportan información útil para estimar la calificación alta.
- La metodología es reproducible y escalable para otros dominios de productos.

La regresión Logística ha posibilitado cuantificar la magnitud y dirección de la influencia de cada factor, logrando así orientar el trabajo a una problemática de gestión de riesgos y oportunidades. La pregunta que se puede ensayar frente a esto sería: ¿Cuál es la probabilidad de que mi producto logre el **status** de alta calificación?

Entendemos que el análisis interpretativo tiene una fortaleza sólida para explicar los impulsores del éxito, pero a su vez comprendemos que como herramienta predictiva puede completarse su estudio debido a la alta varianza no explicada, cuestión que creemos puede ser resuelta a partir de la aplicación de técnicas de *clustering* o la inclusión de variables de *features* específicas de categorías.