Books to Scrape

30 de septiembre

Ciencia de Datos

Proyecto Final

UNIVERSIDAD
Panamericana

Equipo 4:

Ximena Denise Macías Mateos - 0245555

Yael Emiliano Salinas Lozano - 0219039

Edgar Daniel Chacón Amaro - 0213679

Resumen Ejecutivo

El análisis de 1,000 libros indica que con las características disponibles actualmente (Rating, Disponibilidad, Categoría/Genre y presencia de Descripción) no es posible predecir con confiabilidad el precio ni clasificar un libro como “caro”. La evidencia estadística y de modelado sugiere que variables clave de negocio (autor/marca, editorial, formato, año de publicación, serie/franquicia, páginas, señales de demanda/ventas) no están presentes y probablemente explican gran parte de la variación de precio.

Fase de Negocio y Planificación Analítica

Entendimiento del Negocio

Atendemos a la librería "*El Faro*", una librería independiente que cuenta con una tienda física y presencia online en crecimiento. Durante años, la compra de nuevo inventario se ha basado en la intuición del dueño y en las solicitudes directas de los clientes habituales. Sin embargo, tras hacer el análisis, definimos dos problemáticas:

- Exceso de Inventario: Algunos de los libros no se venden y ocupan espacio dentro del almacén, siento este capital inmovilizado.
- Pérdida de ventas: Frecuentemente se agotan títulos con alta demanda inesperada, perdiendo oportunidades de venta y la retención de clientes potenciales.

Objetivo del Proyecto

El objetivo del proyecto es abandonar el modelo de negocio y compra "basado en la intuición" y adoptar una estrategia de reinversión de inventario basada en datos. Buscamos crear un sistema que nos ayude a identificar qué libros tienen la mayor probabilidad de ser populares y rentables para el público objetivo de la librería. Esto servirá como base para optimizar el presupuesto de compras, maximizar la rotación de inventarios y aumentar la rentabilidad a mediano plazo.

Enfoque Analítico

Para lograr nuestro objetivo, definimos un perfil ideal para la librería *El Faro*, analizando cómo interactúan variables clave como el género, el precio, la disponibilidad y la calificación.

- Definición de un Público Objetivo:
    - Nos enfocaremos en jóvenes adultos y adultos contemporáneos (25 - 45 años). Este segmento es activo en redes sociales (influenciados por tendencias como #BookTok), dónde se consume una mezcla de

ficción (fantasía, thrillers, romance) y no ficción (desarrollo, divulgación científica).
    - Ticket Promedio: Buscaremos libros en un rango de precios entre *$30 - $40*. Este rango evita los libros de bolsillo y las ediciones de coleccionista de alto costo y baja rotación.

- Aplicación Práctica
    - Toma de decisiones: Priorizar la compra de libros que se ajusten al perfil ideal (alta calificación, género popular, precio adecuado).
    - Organización (Visuales): Destacar novedades o libros con mayor potencial en lugares estratégicos.

Se definieron las siguientes tareas analíticas principales:

- Predecir: Utilizar modelos de regresión para intentar predecir el precio exacto de un libro. El éxito en esta tarea permitiría a la librería estimar costos y optimizar su presupuesto de compra.
- Clasificar: Implementar modelos de clasificación para segmentar el catálogo en dos categorías: "caro" y "no caro". Esta tarea buscaba crear un mecanismo de filtro rápido para alinear el inventario con el ticket promedio objetivo del negocio.
- Encontrar Relaciones: Antes del modelado, se realizó un análisis estadístico para encontrar relaciones significativas entre las variables.

Estos requerimientos se vieron directamente limitados por la disponibilidad y calidad de los datos. El análisis exploratorio reveló que el dataset carecía de variables críticas (como autor y editorial), lo que impidió que los modelos de predicción y clasificación alcanzaran un rendimiento útil para el negocio y obligó a redefinir los objetivos del proyecto.

Fase de Datos

Extracción de Datos

El dataset se construyó mediante un proceso de Web Scraping automatizado utilizando la librería Selenium. La extracción se ejecutó en dos fases principales para garantizar la integridad y exhaustividad del catálogo:

1. Recolección de URLs: Se utilizó Selenium para simular la navegación web y recolectar de forma recursiva las URLs de detalle de cada libro del catálogo. El proceso implicó la iteración a través de múltiples páginas, resultando en la identificación de 1,000 URLs únicas.
2. Extracción de Atributos: Una vez obtenidas todas las URLs, se visitó cada página de detalle para extraer las siguientes variables clave:
    a. Variables Numéricas Clave: Se extrajo el Precio (sin impuestos) y la Disponibilidad (Stock), junto con el Rating (calificación en estrellas, mapeada a un valor numérico de 1 a 5).
    b. Variables Categóricas/Contextuales: Se recolectaron el Nombre del libro, el Código UPC, la Categoría y la Descripción.
    c. Gestión de Calidad: Se implementó el manejo de excepciones para registrar la presencia o ausencia del texto en el campo Description, generando una feature de control.

El resultado fue un Data Frame inicial que sirvió como base para la posterior limpieza y análisis.

Entendimiento de los Datos (Análisis Exploratorio - EDA)

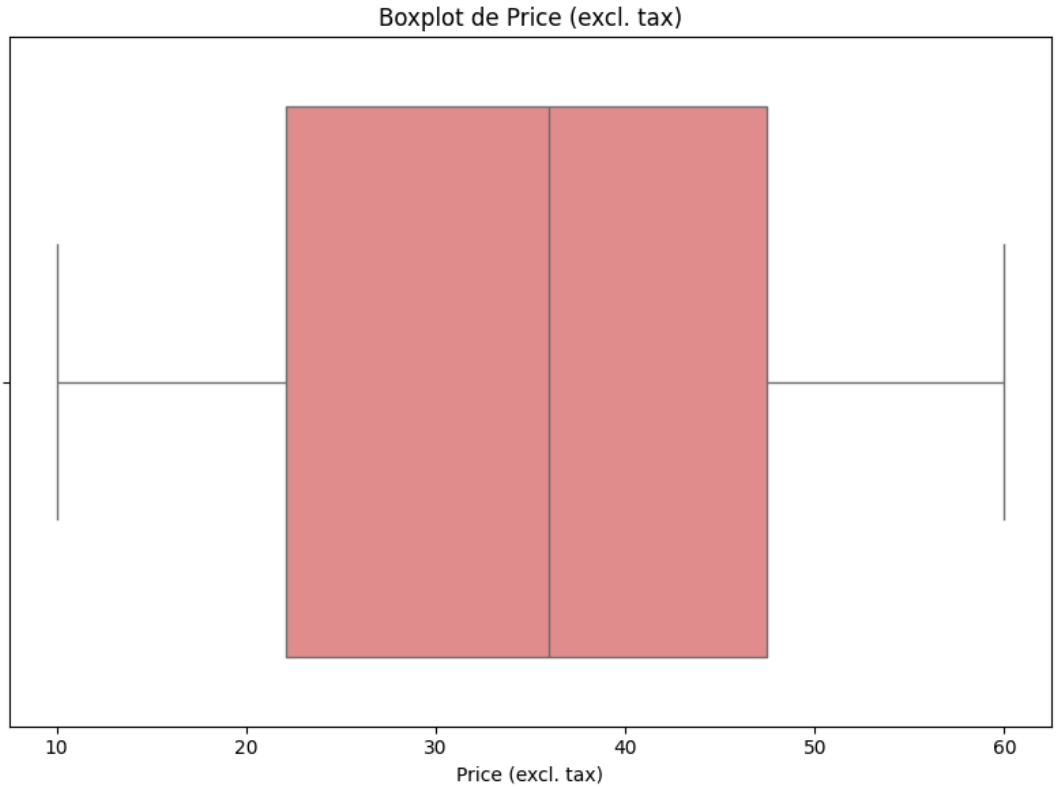
El análisis exploratorio y descriptivo se enfocó en perfilar el catálogo de libros, establecer las bases estadísticas y determinar las relaciones iniciales entre las variables clave (Precio, Rating, y Categoría).

Hallazgos clave

El dataset de 1,000 libros muestra una distribución amplia y con varianza significativa en la mayoría de sus features:

Métrica	Price	Availability	Rating
Media	35.07	8.58	2.92
Mediana	35.98	7	3
Rango (Min - Máx)	10.00 - 59.99	1 - 22	1 - 5
Desviación Estándar	14.45	5.65	1.43

- Precio: El precio promedio es de £35.07, con una alta dispersión (Std=14.45) que indica una gran variación en el valor de los libros. Esta variación es la que los modelos de Machine Learning intentaron, sin éxito, explicar.



- Outliers: No se detectaron outliers significativos en estas tres variables clave, confirmando que la dispersión observada es propia de la distribución de los datos, y no un error.

El Análisis de Componentes Principales se ejecutó para determinar si la alta cantidad de features (54 variables numéricas y codificadas) podía simplificarse.

- El PCA determinó que se necesitan 42 componentes de las 54 variables originales para retener el 80% de la varianza total.

Esta reducción es marginal (solo se eliminan 12 features), lo que conlleva una conclusión crítica para el negocio:

- El problema no es la redundancia: El PCA confirmó que la mayoría de las features (42 de 54) son únicas y aportan varianza.
- Insuficiencia de Datos: El hecho de que estas features únicas no lograran predecir el precio ni clasificar la variable "Caro" (como se vio en los

modelos) significa que ninguna de ellas es relevante para el precio. El modelo simplemente no tiene las variables correctas para la tarea.

Pruebas de Hipótesis

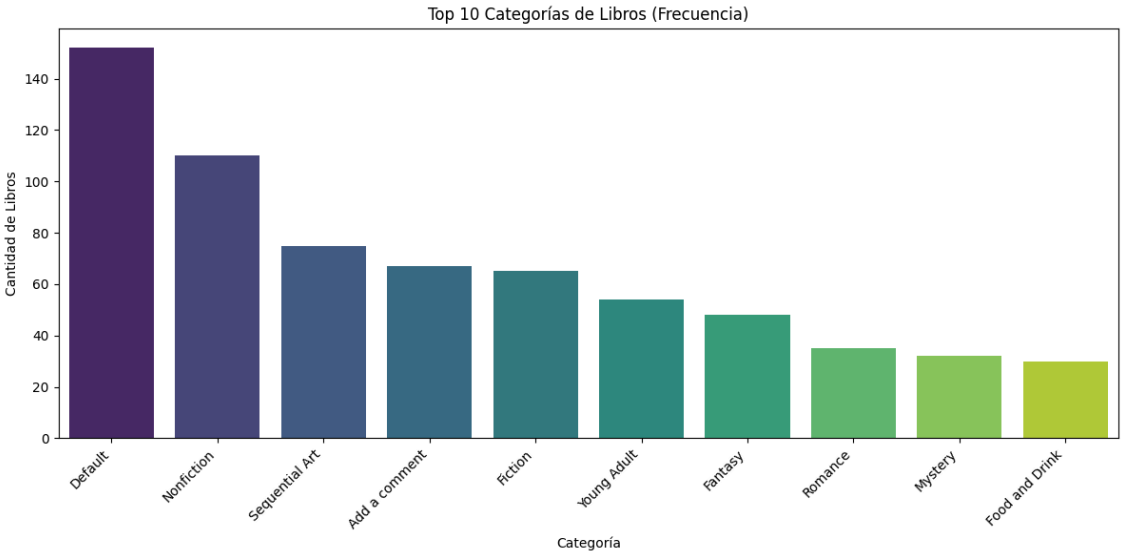
Se plantearon 5 hipótesis para determinar si existían relaciones estadísticamente significativas entre el precio, la calificación, la categoría y la disponibilidad de los libros.

El resultado fue concluyente y consistente en todos los casos: ninguna de las hipótesis nulas fue rechazada. Esto indica que, con los datos disponibles, no se encontró evidencia estadística para afirmar las relaciones que se presuponen; las diferencias observadas en los promedios se deben probablemente al azar y no a un patrón real.

- Hipótesis 1: ¿Los libros con alta calificación (>= 4) son más caros?
    - Resultado: No se encontró evidencia estadística de que los libros con mayor calificación sean más caros (p-value = 0.1404).
- Hipótesis 2: ¿Hay diferencia de precio entre Ficción y No Ficción?
    - Resultado: No se encontró una diferencia de precio significativa entre estas dos grandes categorías (p-value = 0.4320).
- Hipótesis 3: ¿Difieren las calificaciones entre Ciencia Ficción y Misterio?
    - Resultado: No se encontró una diferencia significativa en la calificación promedio entre ambos géneros (p-value = 0.1251).
- Hipótesis 4: ¿Los libros Clásicos son más baratos que el promedio?
    - Resultado: No se pudo confirmar que el precio de los libros clásicos sea significativamente menor al promedio general (p-value = 0.6797).
- Hipótesis 5: ¿Hay diferencia de disponibilidad entre Romance y Thriller?
    - Resultado: No se encontró una diferencia de stock significativa entre estas categorías (p-value = 0.1763).

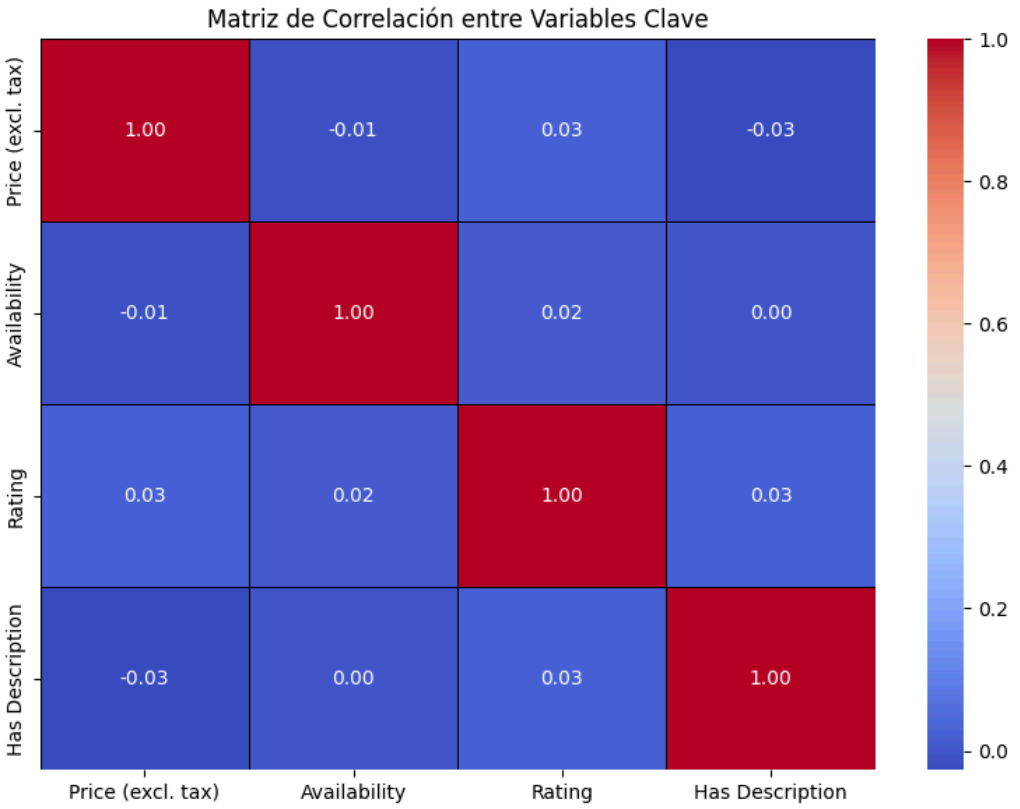
Distribuciones y frecuencias

El análisis de frecuencias muestra el desbalance natural y los problemas de calidad de datos inherentes al catálogo:

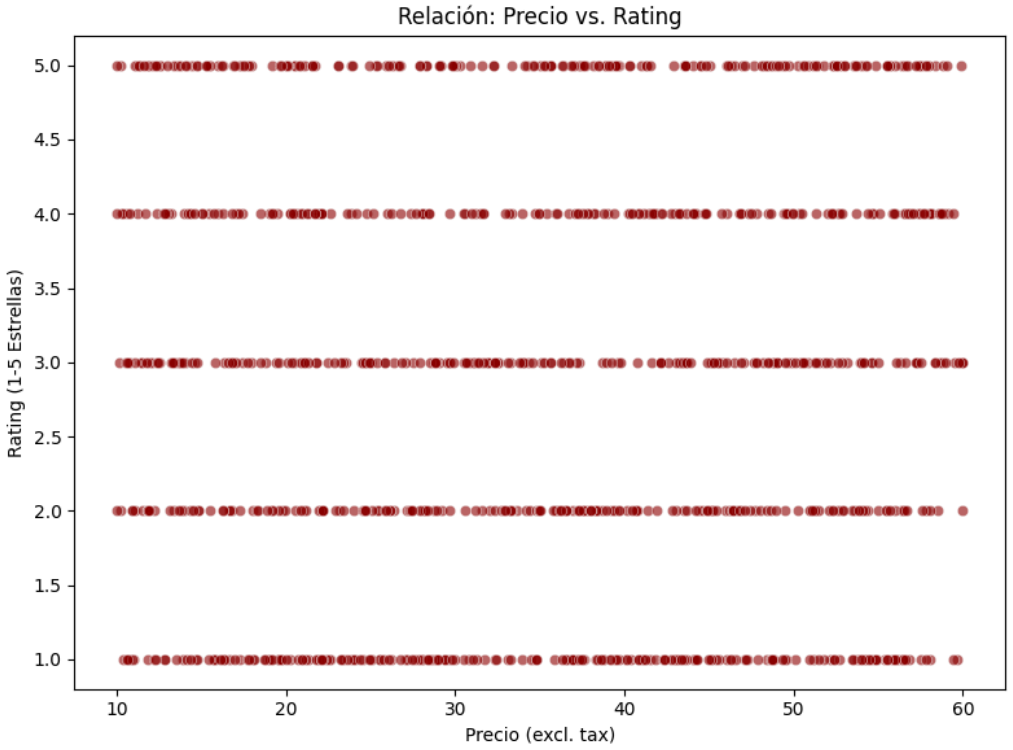


- Desbalance de Categorías: La categoría más frecuente es Default (152 libros), seguida por Nonfiction (110) y Sequential Art (75).
- Problema de Calidad: La categoría "Add a comment" (67 libros) se ubica consistentemente en el Top 5 de frecuencia. Esta etiqueta representa un error en la clasificación, lo cual inyecta ruido al análisis de categorías.

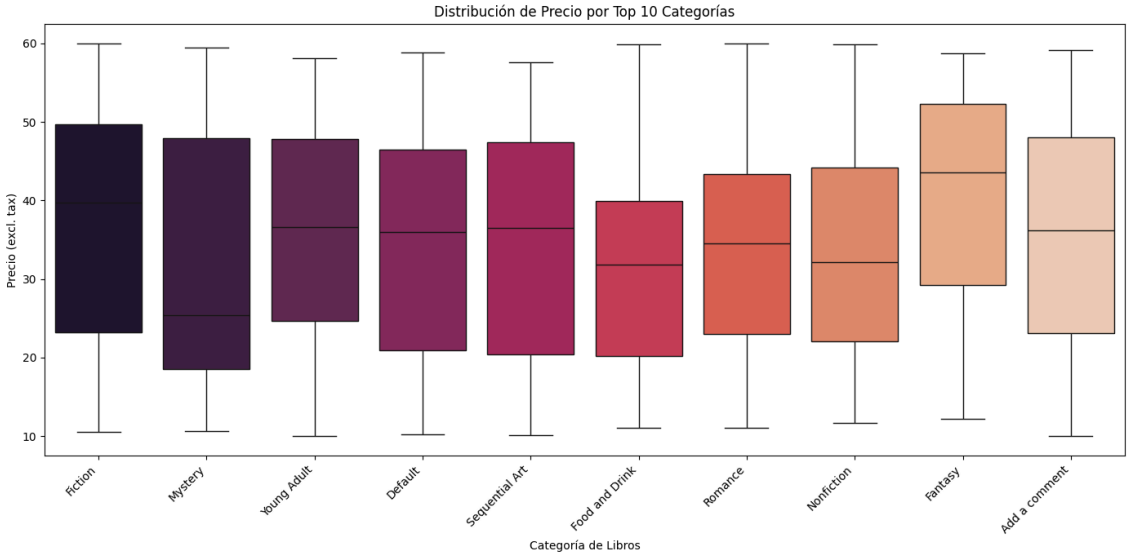
La matriz de correlación demostró que no existe una relación lineal significativa entre las variables clave.



- Precio y Rating (0.03): La correlación entre el precio y la calificación por estrellas es casi nula. Esto desmiente la suposición común de que los libros con mejor Rating son más caros.



El gráfico de precios por las 10 categorías más frecuentes visualmente confirma la falta de distinción entre géneros.



- La mediana del precio se mantiene en un rango estrecho (entre £30 y £40) para casi todas las categorías, con outliers de precio máximo extendiéndose de manera uniforme en la mayoría de los géneros. Esto demuestra que ninguna categoría sobresale claramente como significativamente más cara que las otras, validando los resultados de la prueba de hipótesis (H2: Ficción vs. No Ficción).

Preparación de Datos

La fase de preparación de datos se enfocó en asegurar la calidad, uniformidad y el formato numérico de las variables del dataset (1,000 registros y 13 columnas iniciales).

Proceso de limpieza

Se realizó una vista preliminar para identificar valores nulos y duplicados. Solo se encontraron 2 valores nulos en la columna Description. Dado que la proporción era mínima (0.2% del total) y el campo es contextual, se optó por rellenar estos valores con la etiqueta 'No Description', asegurando que no se perdiera información de otros registros. Adicionalmente, se confirmó que no existían registros duplicados en el dataset.

Posteriormente, se eliminaron las columnas que no aportaban valor predictivo (varianza cero): Product Type, Tax y # Reviews. Esta decisión se justificó porque Product Type mostraba el mismo valor ('Books') para todos los registros , mientras que Tax y # Reviews mostraban un valor de £0.00 o 0 en todos los casos, lo que las hacía inútiles para distinguir patrones en los datos.

Finalmente, se realizó un análisis de Outliers en las tres variables numéricas clave (Price (excl. tax), Rating y Availability) utilizando el método del Rango Intercuartil (IQR). Se determinó que no existen outliers en estas columnas.

Transformaciones

Para estandarizar las features y prepararlas para el modelado, se realizaron las siguientes transformaciones:

- Uniformidad y Tipo de Dato Numérico: Las columnas de precio (Price (excl. tax) y Price (incl. tax)) se recolectaron inicialmente como texto (object) debido a que contenían el carácter de moneda '£'. Se reemplazó este carácter por un vacío y se modificó el tipo de dato a flotante, lo cual es fundamental para el análisis estadístico y la tarea de Regresión.

- Extracción de la Disponibilidad: La columna Availability se transformó extrayendo el valor numérico del stock desde la cadena de texto (ej. "In stock (22 available)"), lo que permitió convertirla a un tipo de dato entero. Esta nueva columna se utilizó como feature numérica.

- Codificación de Variables Categóricas (One-Hot Encoding): La columna categórica Category se transformó utilizando One-Hot Encoding. Esta técnica es necesaria para convertir las etiquetas nominales (como 'Fiction', 'Poetry', etc.) en un conjunto de variables binarias que los algoritmos de Machine Learning pueden procesar. Este proceso resultó en un dataset final de 58 columnas.

Fase de Modelado y Evaluación

Dentro de las preguntas establecidas para el negocio en relación al dataset obtenido se escogieron las siguientes:

- ¿Cuál será el precio (sin impuestos) de un libro en función de su rating, disponibilidad y categoría?
- ¿Un libro será caro (precio por encima de la mediana) o será no caro?

Construcción del Modelo

Dentro de los modelos que se evaluaron de Machine Learning fueron dos en particular con diferentes metodologías (regresión para el caso de la predicción de precio y clasificación para el caso de la etiqueta, es decir, caro o no caro), dentro de las diferentes metodologías utilizadas podemos encontrar:

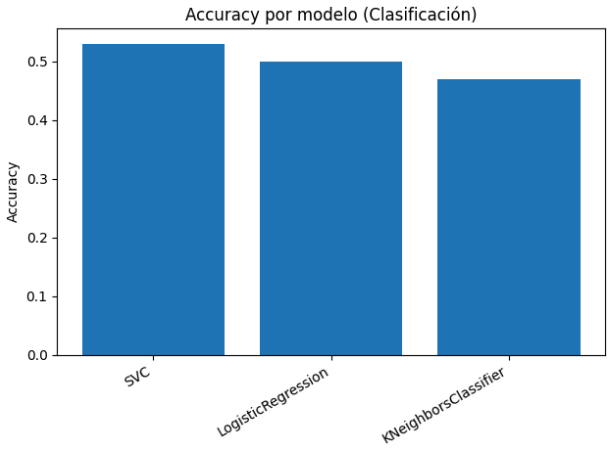
- Regresión Lineal Múltiple
- Regresión Ridge
- Regresión Logística
- Máquina de Vectores de Soporte (SVC)
- K-Vecinos más Cercanos (KNN)

Se escogieron ciertas métricas para la evaluación y relevancia de los modelos implementados con el objetivo de comparar el desempeño que tienen los diferentes modelos:

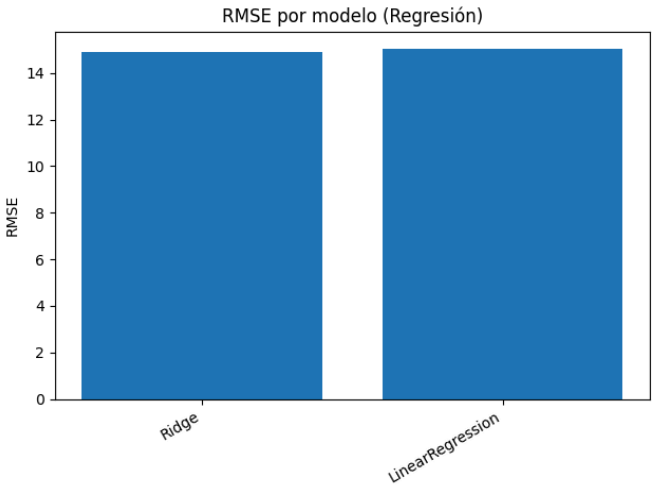
- **Accuracy:** para medir el % de aciertos globales del modelo, es decir, de la totalidad de libros evaluados, definir la proporción de aquellos que fueron clasificados de manera correcta como caro o no caro.
- **RMSE:** esta métrica fue escogida para poder evaluar la precisión de las predicciones del precio. Es calculada a partir de la raíz cuadrada del promedio de los errores al cuadrado (MSE).

Se consideran otras métricas secundarias para verificar otras consideraciones como la estabilidad del modelo y riesgo de sobreajuste, de esta manera comparando las soluciones podemos asegurar que el modelo seleccionado pueda tener buenos números en pruebas históricas y también pueda generalizar bien en el futuro.

Evaluación del Modelo



De acuerdo al desempeño en la clasificación de los tres modelos (Logística, SVC y KNN) estos obtuvieron resultados similares, dando a entender que el predecir la etiqueta “caro” o “no caro” a partir de la variables consideradas es un desafío considerable. Sin embargo, el modelo SVC destacó ligeramente sobre los demás modelos en términos de precisión. En cuanto a la validación el modelo SVC obtuvo un 51% de presión superando a la Regresión Logística con un 49% de precisión y al KNN igualmente con un 49% de precisión. En este sentido, el modelo SVC logró acertar prácticamente la mitad de las veces al identificar si es que un libro pertenece al segmento “caro” o “no caro”. La mejora del modelo SVC respecto a los demás modelos es modesto y a pesar de que el modelo logró capturar patrones leves que los otros modelos no lograron capturar, a partir de estos resultados podemos considerar que las características disponibles (rating, disponibilidad, categoría) no son capaces por sí solas de diferenciar fuertemente a los libros “caros” de “no caros”.



Ahora para el caso del Desempeño en Regresión para predecir el precio exacto de cada libro. En el caso del modelo Ridge fue el que logró dar las estimaciones más precisas. Este modelo tuvo un RMSE promedio de 14 a 15 unidades monetarias, siendo un poco menor (mejor) que el de la regresión lineal tradicional. Este error en términos prácticos significa que en promedio las predicciones de Ridge difieren del precio real en unos $14 (tomando como referencia la moneda oficial del sitio). En comparación con la regresión lineal, esta tuvo un error similar pero con mucho mayor volatilidad con las pruebas y con cierta tendencia a sobre ajustar los datos. Inclusive este tipo de regresión lineal tradicional dio resultados inestables por la presencia de muchas categorías de libros que eran sumamente específicas. El modelo Ridge dio una mayor consistencia con el error y se mantuvo más bajo, debido al mecanismo que se implementa para regularizar y en este sentido penalizar los coeficientes extremos.

De acuerdo a las métricas seleccionadas de desempeño y la confiabilidad consultada, obtenemos que los mejores modelos fueron el SVC y Ridge para las preguntas planteadas en un inicio. Sin embargo, a pesar de las métricas de acierto y la exactitud que obtienen los modelos de predicción están lejos de ser perfectas, pero brindan un punto de partida importante para la toma de decisiones para un alto nivel de dirección.

Para el uso del modelo SVC se puede emplear para poder segmentar el catálogo de libros, de esta manera poder identificar nuevos libros o existentes que puedan caer en la categoría de “caros” y de esta manera ayudar al área de Marketing y Ventas a definir una estrategia diferenciada: con promociones enfocadas, dar una mayor visibilidad a los libros considerados premium o contar con algún tipo de fidelización para clientes que compran libros “caros”.

Para el modelo de regresión Ridge se puede utilizar para simulaciones y estimaciones de precio, que de acuerdo a las características del libro, ya sea género, rating de reseñas, disponibilidad esperada, etc. El modelo puede predecir cuál sería el precio probable del libro y con esta predicción ayudar a fijar precios de nuevos lanzamientos de libros o ajustarlos con el objetivo de competir de la mejor manera en el mercado. También es posible plantear casos hipotéticos en cuanto a los precios y observar de qué manera el hecho de que un libro pertenezca a una categoría distinta o reciba un mejor puntaje en su calificación pueda afectar en el precio final del libro, lo cual puede ayudar a explorar diversos escenarios para el negocio.

Fase de Despliegue y Conclusiones

Propuesta de Despliegue

Dado que el principal hallazgo del proyecto fue la insuficiencia de los datos, la propuesta de despliegue se enfoca en un enfoque iterativo y realista. En lugar de entregar un modelo predictivo fallido, entregamos valor a través de la inteligencia de negocio generada.

- Fase 1: Informe de Diagnóstico → Entrega Inmediata
    - Su principal valor es estratégico. Previene que la librería invierta recursos en una estrategia de datos basada en variables incorrectas. El informe redefine el problema y presenta un caso de negocio claro para la recomendación principal: enriquecer el dataset con variables de alto impacto (autor, editorial, ventas, fechas).
- Fase 2: Dashboard Interactivo de Catálogo → Corto / Mediano Plazo
    - Un dashboard interactivo conectado al nuevo dataset. El objetivo será recaudar nuevos datos que vayan acorde a un modelo mejorado que entregue valor.
- Fase 3: Sistema de Soporte a la Decisión de Compra → Largo plazo
    - Visión final del proyecto; un programa que brinde soporte en cada compra de inventario (mensualmente) que logré puntuar qué libros son prioridad de compra, este ya estaría alineado con los objetivos originales.

Conclusiones y Retroalimentación

Hallazgos principales

Los hallazgos son concluyentes y establecen que el dataset actual es insuficiente para el objetivo predictivo:

- Fallo Estadístico Clave: Las pruebas de hipótesis no encontraron una diferencia de precio estadísticamente significativa basada en el Rating o la Categoría. La correlación entre el precio y el Rating es casi nula (0.03), desmintiendo la suposición de que los libros mejor calificados son más caros.

- Modelo sin Valor Predictivo: Los modelos de Machine Learning fallaron en aprender a predecir o clasificar. El error de predicción (RMSE del Ridge) fue similar a la variación natural de los precios (Std=£14.45), y la Accuracy para clasificar un libro como "Caro" (SVC) fue de tan solo 53%.
- Justificación de Falla (PCA): El PCA confirmó que la falla no es por redundancia, sino por la ausencia de las variables correctas. El precio está gobernado por factores no capturados que determinan la popularidad y rentabilidad.

Limitaciones del modelo

- Insuficiencia de Features: Las variables recolectadas son inadecuadas para predecir la rentabilidad y popularidad.
- Ausencia de Drivers de Precio: Las features críticas que impulsan el valor del libro en el mercado (como Autor, Editorial, o Año de Publicación) no se incluyeron en el scrapping inicial. Esta es la principal limitación que impide mitigar la Pérdida de Ventas por demanda inesperada.

Recomendaciones Estratégicas

El camino a seguir debe enfocarse en el enriquecimiento del dataset para construir modelos con valor predictivo y cumplir con el objetivo de abandonar el modelo basado en la intuición:

- Prioridad: Enriquecimiento de Datos: Es el paso más crucial. Modificar la estrategia de scrapping para obtener urgentemente las features de alto valor predictivo: Autor y Editorial.
- Limpieza de Texto Avanzada (NLP): Utilizar el campo Description para análisis de texto (NLP) y extraer features estructuradas (ej., longitud o sentimiento).
- Modelado Avanzado (Random Forest/XGBoost): Solo después de la fase de enriquecimiento, se emplearán algoritmos más robustos para crear el sistema basado en datos que permitirá a la librería "El Faro" tomar decisiones estratégicas.