Andrés Mauricio Plazas González

PREICA2501B020071

Programa Académico:

Ingeniería De Software y Datos

Actividades:

S25 - Evidencia de Aprendizaje 1 - Proyecto Integrado III

Asignatura :

Proyecto Integrado III

Docente:

Sharon Karin Camacho

Institución Universidad Digital de Antioquia

10 de mayo del 2025

Introducción

En el dinámico mundo del comercio electrónico, la vasta cantidad de datos generados por las interacciones de los usuarios representa una fuente invaluable de conocimiento para la toma de decisiones estratégicas. Esta actividad se centra en el análisis exploratorio de un rico dataset de comportamiento de comercio electrónico, disponible en Kaggle (enlace al dataset). El objetivo principal es iniciar un proceso de análisis riguroso que nos permita identificar problemas de negocio relevantes, comprender la estructura y las características de los datos, y sentar las bases para futuras investigaciones y modelos predictivos.

I. Definición del problema de negocio

- **Descripción clara y concisa del problema a abordar:**

  La falta de comprensión detallada de los patrones de comportamiento de los usuarios en las diferentes etapas de su interacción con la plataforma de comercio electrónico dificulta la optimización de estrategias de marketing, la personalización de la experiencia del usuario y la identificación de oportunidades para aumentar las tasas de conversión y el valor promedio del pedido.

  Por lo cual existe la necesidad de obtener una comprensión profunda del comportamiento de los usuarios en el sitio web de comercio electrónico para identificar patrones que puedan informar estrategias destinadas a mejorar la conversión y la experiencia del usuario.

- **Formulación de una pregunta de investigación específica y medible:**

  En el competitivo panorama del comercio electrónico, comprender a fondo el comportamiento de los usuarios en línea es esencial para impulsar el crecimiento y optimizar la experiencia de compra. Por lo cual daremos respuesta a la siguiente pregunta *"¿Cómo varía el comportamiento de los usuarios a lo largo del tiempo (analizando las tendencias diarias y semanales) y entre diferentes categorías de productos, patrones temporales o por categoría que se correlacionan con una mayor tasa de conversión dentro de una misma sesión de usuario?"* a lo largo de la asignatura.

- **Definición de métricas para evaluar el éxito del análisis:**

  El éxito de este análisis se evaluará en función de nuestra capacidad para extraer insights accionables que puedan impactar positivamente el negocio de comercio electrónico. A muy alto nivel, las métricas clave se centran en la comprensión del comportamiento del usuario y la identificación de oportunidades para la mejora.

  A continuación, se describen las métricas principales de forma resumida pero detallada:

  1. **Profundidad del Insight sobre el Comportamiento del Usuario:**

     ¿Qué mide? La riqueza y la granularidad de los hallazgos sobre cómo los usuarios interactúan con la plataforma. Esto incluye la identificación de patrones temporales (diarios, semanales), diferencias en el comportamiento entre categorías de productos y la comprensión de las secuencias típicas de eventos que conducen o no a una compra.

     ¿Cómo se evalúa? Se valorará la capacidad del análisis para ir más allá de las estadísticas descriptivas básicas y revelar relaciones y tendencias significativas. Por ejemplo, identificar horas pico de actividad por categoría, secuencias de navegación comunes antes del abandono del carrito, o categorías con altas tasas de visualización pero bajas tasas de compra.

2. **Identificación de Correlaciones con la Tasa de Conversión:**

   ¿Qué mide? La habilidad para encontrar patrones específicos de comportamiento (temporales o por categoría) que se asocian con una mayor probabilidad de que un usuario complete una compra.

   ¿Cómo se evalúa? Se considerará exitoso el análisis si logra identificar momentos específicos (días, horas) o categorías de productos donde la secuencia "view" -> "cart" -> "purchase" es más frecuente. Esto podría implicar el cálculo de tasas de conversión condicionadas a ciertos comportamientos o momentos.

3. **Potencial de Accionabilidad de los Hallazgos:**

   ¿Qué mide? La relevancia práctica de los insights generados para la toma de decisiones de negocio. ¿Los hallazgos sugieren acciones concretas que la empresa puede implementar?

   ¿Cómo se evalúa? Se valorará si las conclusiones del análisis pueden traducirse en recomendaciones específicas para optimizar campañas de marketing (ej., enfocar esfuerzos en ciertas horas o categorías), mejorar la experiencia del usuario (ej., identificar puntos de fricción en el proceso de compra en categorías específicas) o personalizar ofertas.

4. **Claridad y Solidez de las Conclusiones:**

¿Qué mide? La coherencia, la lógica y el respaldo de los hallazgos del análisis con los datos explorados. ¿Las conclusiones están bien fundamentadas y son fáciles de entender?

¿Cómo se evalúa? Se revisará la claridad con la que se presentan los resultados y la robustez de las evidencias que los respaldan, basándose en la exploración realizada con Pandas Profiling y el análisis posterior.

II. **Fuentes de datos:**El dataset que analizaremos en esta actividad proviene de una recopilación realizada por el proyecto Open CDP, y encapsula un mes completo de interacciones de usuarios (Octubre de 2019) en una extensa tienda en línea multicategoría. Este conjunto de datos se presenta como una valiosa ventana al comportamiento digital de los consumidores dentro de un entorno de comercio electrónico real y diverso.

Enlace dataset:

https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

III. **Planeación:** Para la gestión eficiente y colaborativa de este proyecto de análisis del dataset de comportamiento de comercio electrónico, se ha creado un tablero dedicado en la plataforma Trello. Este entorno virtual de trabajo permitirá al

equipo visualizar el progreso de las tareas, asignar responsabilidades y mantener un seguimiento claro de los plazos establecidos.

Enlace al tablero de Trello:

https://trello.com/invite/b/681fd453e90597886fc94ae0/ATTIcbf3ce9c626eb99f4bb2ff23b9fc74c4CE629C04/proyecto-integrado-iii-analitica-de-datos-sharon-karin-camacho-preica2501b020071

IV. Creación del repositorio:

https://github.com/AndresPlazas19931504/PROYECTOINTEGRADOR-PREICA2501B020071

V. Utiliza *pandas_profiling* **para explorar el dataset. Incluye observaciones sobre:**

- Duplicados.
- Valores nulos.
- Tipos de datos.
- Distribuciones.
- Correlaciones.
- Resumen estadístico.

En nuestro análisis del dataset de comportamiento de comercio electrónico, hemos llevado a cabo una exploración inicial exhaustiva para comprender su

estructura, calidad y las características clave de la información que contiene. A continuación, detallamos los pasos principales que hemos realizado:

1. Selección e Importación de Herramientas (Bibliotecas de Código):

Se seleccionan las bibliotecas necesarias para este análisis, nosotros importamos bibliotecas de código especializadas en el manejo y análisis de datos. por lo cual utilizamos:

- **Pandas:** Esta biblioteca es fundamental para trabajar con datos estructurados en forma de tablas, similar a hojas de cálculo avanzadas. Nos permite organizar, manipular y analizar la información de manera eficiente.
- **Kagglehub:** Dado que el dataset se encuentra alojado en la plataforma Kaggle, utilizamos kagglehub para facilitar la descarga directa y segura del archivo de datos a nuestro entorno de trabajo.
- **ydata-profiling (anteriormente pandas-profiling):** Esta potente herramienta nos permite generar un reporte de análisis exploratorio de datos de forma automatizada. Es como tener un analista experto que realiza un primer vistazo detallado al dataset y resume sus hallazgos clave.

2. Descarga e Ingreso de los Datos:

Nuestro siguiente paso fue obtener el corazón de nuestro análisis. Utilizando kagglehub, descargamos el archivo específico del dataset de comportamiento de comercio electrónico correspondiente al mes de octubre de 2019 (2019-Oct.csv). Una vez descargado, cargamos este archivo en una estructura de datos de Pandas llamada

"DataFrame" (df). Este DataFrame como una tabla gigante donde cada fila representa un evento individual (una acción de un usuario) y cada columna representa una característica de ese evento (tipo de evento, producto, usuario, etc.).

3. Un Primer Vistazo a la Información ("Ojeando los Datos"):

Realizamos una inspección general del DataFrame para tener una idea de su contenido y formato. Utilizamos comandos como:

- **head():** Para visualizar las primeras filas del DataFrame, lo que nos da una muestra rápida de los tipos de eventos y la información que contienen las columnas.
- **info():** Esta función nos proporcionó un resumen conciso del DataFrame, incluyendo el número total de filas, el nombre de cada columna, el tipo de datos que contiene (por ejemplo, números enteros, texto, fechas) y la cantidad de valores no faltantes en cada columna.
- **describe():** Para las columnas que contienen datos numéricos, describe() nos ofreció un conjunto de estadísticas descriptivas clave, como la cantidad de datos, el promedio, la desviación estándar (una medida de la dispersión), los valores mínimo y máximo, y los percentiles (que nos indican la distribución de los valores).

4. Exploración Detallada de las Columnas:

Luego, profundizamos en la comprensión de cada columna individualmente para conocer los valores específicos que contiene:

- **unique():** Aplicamos esta función a cada columna para identificar todos los valores distintos o únicos que aparecen en ella. Por ejemplo, en la columna "event_type", pudimos ver los diferentes tipos de acciones registradas (como "view", "cart", "purchase").
- **value_counts():** Para cada columna, utilizamos value_counts() para contar la frecuencia con la que aparece cada valor único. Esto nos dio una idea de la distribución de las diferentes categorías dentro de cada columna (por ejemplo, cuántas veces ocurrió cada tipo de evento).

5. Generación de un Reporte de Análisis Automatizado:

Para obtener una visión aún más completa y detallada del dataset de manera eficiente, recurrimos a la herramienta ydata-profiling. Instalamos esta biblioteca adicional que automatiza la generación de un reporte de exploración de datos muy completo.

6. Creación del "Resumen Ejecutivo" Automático:

Finalmente, utilizando ydata-profiling y aplicando el análisis a una muestra representativa del 12.5% de nuestros datos (para optimizar el tiempo de procesamiento sin perder información clave), generamos un archivo HTML llamado reporte.html. Este reporte es como un "resumen ejecutivo" creado automáticamente por la computadora.

Contiene una gran cantidad de información organizada de manera visual e intuitiva, incluyendo:

- **Gráficos de distribución:** Muestra cómo se distribuyen los valores en cada columna.
- **Estadísticas detalladas:** Incluyen medidas de tendencia central, dispersión, y otros indicadores relevantes.
- **Información sobre valores faltantes y duplicados:** Alerta sobre posibles problemas de calidad de los datos.
- **Matrices de correlación:** Ayuda a identificar posibles relaciones entre las diferentes variables numéricas.

Este reporte reporte.html nos proporciona una visión panorámica y detallada del dataset, facilitando la identificación de patrones, anomalías y características importantes que serán fundamentales para las siguientes etapas de nuestro análisis y para responder a nuestra pregunta de investigación.

Enlace:

https://colab.research.google.com/drive/1YiiE0GPjuzP2nm2PiW2JeykhuWzo9zGC?usp=sharing