Andrés Mauricio Plazas González

PREICA2501B020071

Programa Académico:

Ingeniería De Software y Datos

Actividades:

S20 Evidencia de aprendizaje 2 - Limpieza y transformación del conjunto de datos

Asignatura :

Proyecto Integrado III

Docente:

Sharon Karin Camacho

Institución Universidad Digital de Antioquia

27 de mayo del 2025

Introducción

Para asegurar la solidez de nuestros análisis sobre el comportamiento de e-commerce, esta sección se dedica a la mejora de la calidad del dataset. Hemos llevado a cabo una serie de pasos de limpieza y preprocesamiento esenciales. Esto incluyó la eliminación de registros duplicados y el tratamiento exhaustivo de valores nulos, tanto eliminando filas en campos críticos como imputado en otros. También nos enfocamos en la conversión de tipos de datos para columnas clave como event_time, category_code, brand y user_session, asegurando su formato correcto. Finalmente, abordamos los valores atípicos en la columna price utilizando Winsorización y corregimos errores tipográficos en las columnas categóricas para una mayor uniformidad. Todas estas medidas garantizan un dataset limpio y optimizado para análisis más profundos.

Descripción del desarrollo

Basado en el análisis y las acciones realizadas en el notebook después de "S25 - Evidencia de aprendizaje 1", podemos describir las necesidades de limpieza en los siguientes aspectos:

Duplicados: Se identificó la existencia de filas duplicadas en el dataset. Aunque no se especifican las columnas exactas donde se encuentran, la eliminación de duplicados se realizó sobre todas las columnas (df.drop_duplicates(inplace=True)), lo que aborda la necesidad de asegurar que cada evento registrado sea único. El código mostró que se eliminaron una cantidad significativa de filas duplicadas, lo que confirma su presencia inicial.

Valores Nulos: Se identificaron columnas con valores nulos. Específicamente, se encontró que la columna user_session contenía valores nulos que fueron considerados críticos para el análisis de comportamiento del usuario, por lo que las filas con nulos en esta columna fueron eliminadas. Otras columnas con nulos potenciales (como category_code y brand, aunque en el análisis inicial no mostraron nulos después de la carga optimizada, es común que puedan presentarlos en otros contextos o con carga no optimizada) fueron tratadas mediante imputación con la moda. La proporción de datos faltantes en user_session fue significativa antes de la eliminación. La relevancia de tratar estos nulos radica en la necesidad de tener información completa para realizar análisis de sesiones de usuario y entender la actividad asociada a cada evento.

Inconsistencias en Valores: Se revisaron las columnas category_code y brand en busca de inconsistencias o errores tipográficos. Se encontraron y corrigieron problemas como espacios adicionales al final de algunas categorías y marcas ('electronics.smartphone ', 'samsung ', 'apple ', 'xiaomi ') y se mapeó 'electronics.telephone' a 'electronics.smartphone'. Estas correcciones se realizaron utilizando el método replace() y cat.rename_categories() para asegurar la estandarización de los valores categóricos y facilitar análisis y agrupaciones precisas.

Tipos de Datos: Se identificó que la columna event_time no estaba en el formato de fecha y hora adecuado para realizar análisis temporales. La corrección necesaria fue convertir esta columna al tipo de dato datetime de pandas (pd.to_datetime(df['event_time'])). Además, se aseguró que las columnas event_type, category_code, brand y user_session fueran de tipo category para optimizar el uso de memoria y mejorar el rendimiento en operaciones de agrupación y filtrado.

Valores Atípicos: La columna price fue identificada como una variable donde los valores atípicos (outliers) podrían ser problemáticos, especialmente en el extremo superior (precios muy altos) que podrían distorsionar análisis de precios promedio o gasto total. Se consideró necesario un tratamiento especial para estos valores atípicos. La técnica aplicada fue la Winsorización (Winsorizer(capping_method='gaussian', tail='right', fold=1.5, variables=['price'])), que limita los valores extremos a un umbral

definido (en este caso, 1.5 desviaciones estándar por encima de la media), mitigando su impacto sin eliminarlos por completo.

Nivel de Granularidad: El dataset tiene una granularidad a nivel de evento, donde cada fila representa una acción individual de un usuario en un momento específico. Este nivel de detalle es mayor al requerido si solo se necesitan datos agregados a nivel mensual. Aunque la granularidad a nivel de evento es valiosa para análisis detallados (embudos de conversión, secuencias de eventos), se evaluó la necesidad de agregar datos para análisis con granularidad menor. Como se demostró en la sección de "Análisis Exploratorio y Agregaciones", se agregaron datos por mes y categoría para analizar las ventas, lo que confirma que la agregación es necesaria y posible para lograr la granularidad deseada en función del análisis específico.

Enlace dataset limpio:

https://drive.google.com/file/d/1dPtJUmCK139kXWdpWtjwMovRqRHPk5vb/view?usp=sharing

Creación del repositorio:

https://github.com/AndresPlazas19931504/PROYECTOINTEGRADOR-PREICA2501B020071/blob/main/S20_Evidencia_de_Aprendizaje_2.ipynb

Enlace .py:

https://colab.research.google.com/drive/1fttbMjUz41c83qzw1YK7773nGAWMyHAQ?usp=sharing

Conclusión

El dataset limpio parece estar en una buena condición para proceder con los análisis. Basándome en la evaluación de calidad de datos que realizamos, se puede responder a las siguientes preguntas sobre la completitud, relevancia y granularidad del dataset limpio.

Completitud de los Datos: Como se mostró en la sección "Evaluación de la Completitud de los Datos", verificamos las columnas con valores nulos restantes después de todo el tratamiento. Se eliminan filas con valores nulos en la columna user_session y se imputaron valores nulos en otras columnas con la moda. Las columnas críticas como event_time, event_type, product_id, user_id, user_session y price no tienen valores nulos restantes. Esto asegura que los análisis que dependen de estas columnas no se vean afectados por datos faltantes.

Relevancia de las Variables: En la sección "Evaluación de la Relevancia de las Variables", listamos las columnas disponibles (event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session, month). Estas columnas son fundamentales para analizar el comportamiento de e-commerce, los patrones de compra, la popularidad de productos y marcas, y la actividad de los usuarios a lo largo del tiempo. Por lo tanto, las variables presentes son relevantes para responder a los objetivos de análisis iniciales.

Granularidad Adecuada: La granularidad del dataset es a nivel de evento, lo que significa que cada fila representa una acción individual de un usuario (ver, añadir al carrito, comprar). Esta granularidad detallada es ideal para análisis profundos como embudos de conversión y secuencias de eventos. Además, como se demostró en la

sección "Análisis Exploratorio y Agregaciones", podemos agregar los datos a niveles menos detallados (por ejemplo, ventas mensuales por categoría) para obtener insights agregados cuando sea necesario. La granularidad a nivel de evento nos da la flexibilidad de realizar análisis tanto detallados como agregados.

En conclusión, basándonos en la evaluación realizada, el dataset limpio parece adecuado para proceder con los análisis y abordar los objetivos de negocio iniciales, ya que la completitud, relevancia y granularidad son apropiadas.