Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial Aplicada

Materia:

Análisis de grandes volúmenes de datos

Actividad:

4.3 Avance de proyecto 1: Sistema de Recomendación

Equipo 40:

Cecilia Acevedo Rodríguez- A01793953

Francisco Xavier Bastidas Moreno - A01794188

Ricardo Mar Cupido – A01795394

Edgar Gerardo Rojas Medina - A00840712

Tabla de contenido

Plan de proyecto y Cronograma

En este documento se presentará el plan de proyecto para el desarrollo de un sistema de recomendación para el comercio electrónico. Se incluirá un cronograma detallado de actividades que guiará el proceso de implementación.

El plan de proyecto se enfoca en la metodología CRISP-DM la cual consiste en 6 fases:

- Entendimiento del negocio: Se define el objetivo del proyecto y las necesidades del proyecto.
- Entendimiento de los datos: Se recopilan y exploran los datos, además, se realiza la identificación de las variables clave.
- Preparación de los datos: En este paso los datos se limpian y sufren una transformación para la etapa de modelado.
- Modelado: Se realiza la selección y entrenamiento con los modelos predictivos utilizando los datos limpios y procesados en los pasos anteriores.
- Evaluación: Se realiza la evaluación del rendimiento del modelo mediante el uso de métricas como precisión, recall y F1-Score. Además, en esta etapa se realiza el ajuste a los modelos para obtener mejores resultados en estas métricas.
- Despliegue: Es la implementación del modelo en el entorno de producción una vez realizado los pasos anteriores.

A continuación, se presenta el cronograma de actividades a realizar a lo largo de este trimestre, basado en la metodología CRISP-DM:

Objetivo

El objetivo de este trabajo es responder la siguiente pregunta: *¿Cuáles son los productos de mayor interés según su historial de calificaciones?* Para abordar esta pregunta, utilizaremos una base de datos de Amazon que incluye diversas calificaciones, productos, categorías y número de clientes. Este dataset se puede encontrar en el siguiente enlace: https://github.com/MengtingWan/marketBias/blob/master/data/df_electronics.csv

Justificación y preprocesamiento

Se utilizó un conjunto de datos de Amazon debido a su conocida variedad de productos y la riqueza de datos sobre las preferencias y comportamientos de los usuarios.

1. **Variedad de productos**: Amazon ofrece una amplia gama de productos en diversas categorías, lo que permite construir un sistema de recomendación que pueda abordar diferentes intereses y necesidades de los usuarios.
2. **Volumen de datos**: Amazon maneja enormes cantidades de datos de transacciones y opiniones de los usuarios, proporcionando una gran cantidad de información para entrenar y evaluar modelos de recomendación.
3. **Diversidad demográfica de usuarios**: Los usuarios de Amazon provienen de diversos ámbitos geográficos y demográficos, lo que garantiza una representación variada en el conjunto de datos y permite construir modelos de recomendación relevantes para una audiencia amplia.

En cuanto a los pasos de preprocesamiento, solo fue necesario seleccionar las columnas relevantes para el sistema de recomendación y realizar una limpieza básica de los datos.

Ejercicio de exploración inicial y análisis del conjunto de datos

Se llevó a cabo un ejercicio de exploración del dataset, comenzando con una descripción del mismo. El conjunto de datos tiene un tamaño de 1,292,954 filas por 10 columnas. Las columnas del dataset son las siguientes:

1. item_id: identificador del producto.
    a. Cantidad de registros: 1292954
    b. Tipo de dato: int64
2. user_id: identificador del usuario.
    a. Cantidad de registros: 1292954
    b. Tipo de dato: int64
3. rating: Calificación del product dada por el usuario
    a. Cantidad de registros: 1292954
    b. Tipo de dato: float64
4. Timestamp: Fecha de registro
    a. Cantidad de registros: 1292954
    b. Tipo de dato: object
5. model_attr:
    a. Cantidad de registros: 1292954
    b. Tipo de dato: object
6. Category: Categoría del producto
    a. Cantidad de registros: 1292954
    b. Tipo de dato: object
7. Brand: Marca del product
    a. Cantidad de registros: 331120
    b. Tipo de dato: object
8. Year: Año en el que se realize el registro
    a. Cantidad de registros: 1292954
    b. Tipo de datos: int64
9. user_attr: Atributos del usuario
    a. Cantidad de registros: 174124

b. Tipo de dato: object

10. Split:
    a. Cantidad de registros: 1292954
    b. Tipo de dato: int64

En este ejercicio, mostramos la distribución del rating de todo el conjunto de datos (Figura 1). A simple vista, se puede observar que la mayoría de las calificaciones son positivas. Para obtener una comprensión más profunda de los datos, mostramos también la distribución de las categorías (Figura 2). Además, mostramos la distribución de las fechas en las que se realizaron las valoraciones de los productos. Como se puede ver en el gráfico de la figura 3, la mayoría de las valoraciones se realizaron entre los años 2015 y 2017.

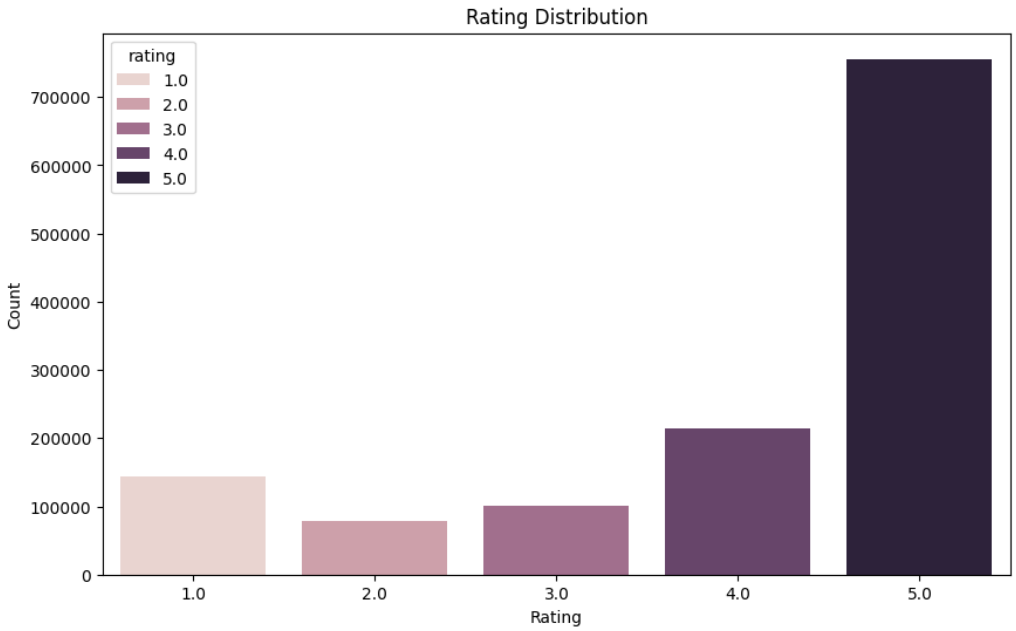

Figura 1. Distribución del rating.

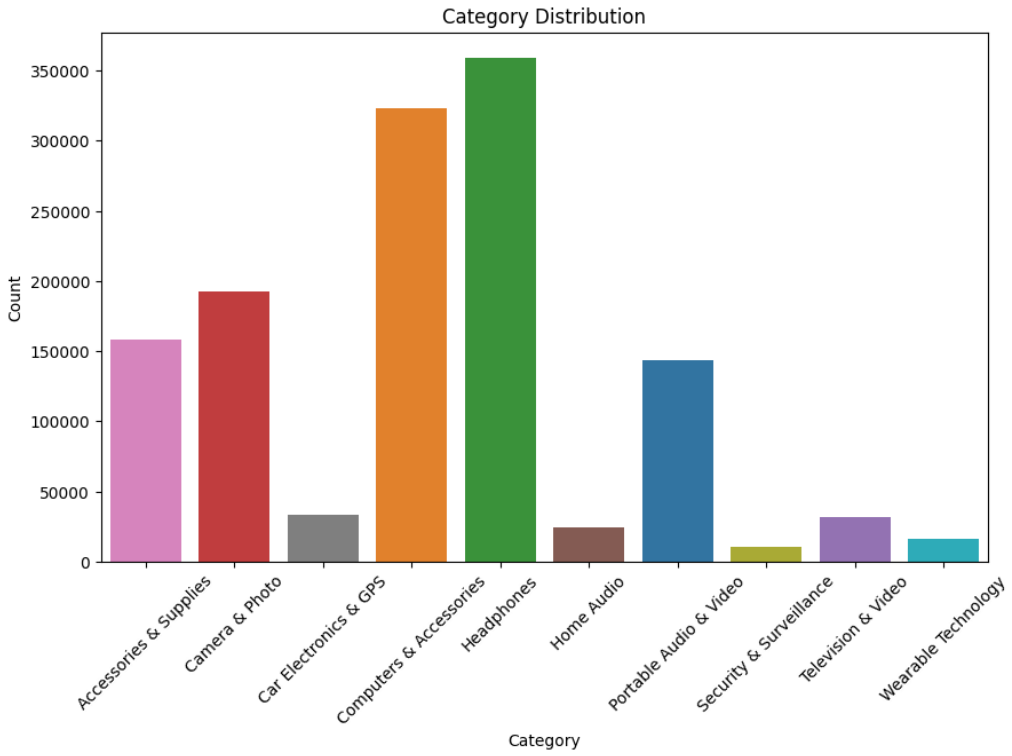

Figura 2. Distribución de categorías.

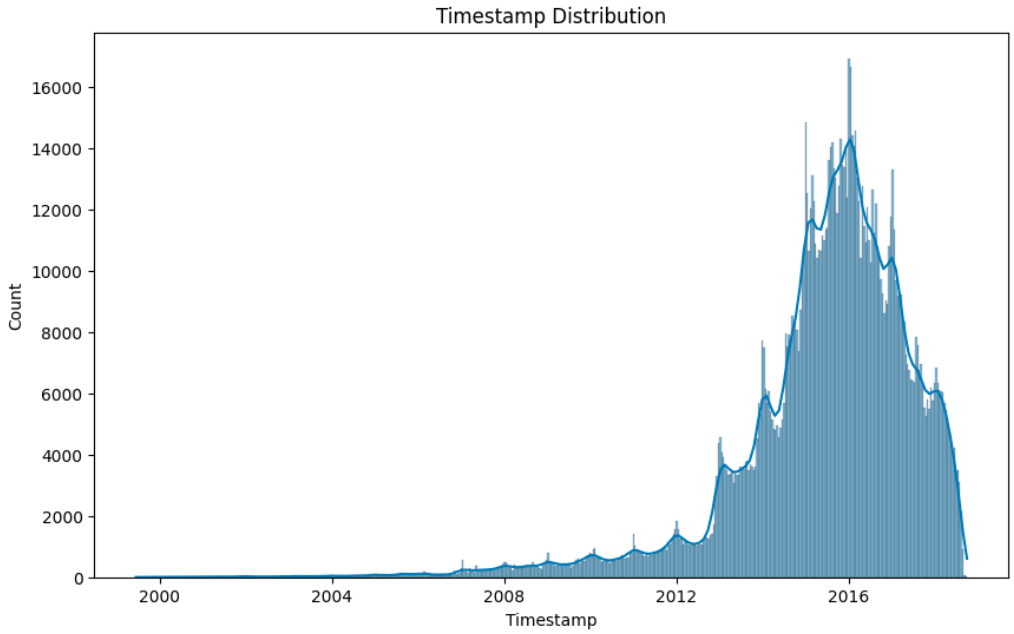

Figura 3. Distribución de las fechas en las que se realizaron las valoraciones de los productos.

Algoritmo de recomendación básico

El algoritmo que desarrollamos un algoritmo de filtros colaborativos, también conocido como recomendación item – item, para esta entrega utilizaremos K-Means, el cual es uno de los algoritmos de clusterización más utilizados en los sistemas de recomendación (Mulyawan et al., 2019b). Este clasificador tendrá los siguientes datos de entrada: user_id, item_id y rating. Esto último con el fin de obtener una lista de productos que puedan ser recomendados en el usuario basado en los “ratings” que el usuario haya realizado en el pasado.

De acuerdo con Mulyawan et al. (2019), el método K-Means divide los datos en grupos, de modo que los datos que tenían las mismas características se agrupan en el mismo grupo y los datos con características diferentes se agruparan en otros grupos. El propósito de la agrupación es ***objetos*** hasta que la distancia de cada objeto al centro del grupo en un grupo sea mínima.
El proceso de agrupación, utilizando K-Means, se lleva a cabo con un algoritmo que sigue el siguiente orden:

1. Se determina el número de clústeres.
2. Se asignan datos en grupos al azar.
3. Se calcula el centroide (promedio) de los datos en cada grupo.
4. Se asigna cada dato al centroide más cercano.
5. Se regresa a la etapa 3, si todavía hay datos que mueven los clústeres o si los cambios en el valor del centroide están por encima del valor umbral especificado.

Resultados

Para la evaluación de nuestro modelo se utilizaron las métricas MAE y RMSE para la evaluación del modelo, de acuerdo con Wang and Lu (2018) tanto la métrica MAE

y RMSE son muy utilizadas para la evaluación de sistemas de recomendación. La diferencia entre el RMSE y el MAE radica en el hecho de que el primero prefiere sistemas que cometen pocos errores, mientras que el segundo prefiere sistemas que cometen una menor cantidad de errores.

Por lo tanto, RMSE es más útil cuando los errores grandes son especialmente indeseables o cuando los errores se distribuyen normalmente. Mientras que MAE es menos susceptible a los valores atípicos y nos proporciona información sobre el error promedio del modelo. Sin embargo, en ambos casos, entre más cercano este el valor de la métrica al cero, más precisión tiene nuestro modelo.

Los resultados de estas 2 métricas fueron los siguientes:

RMSE: 1.3749
MAE: 1.1003

Como puede observarse, existe un margen de mejora para nuestro sistema de recomendación básico, sin embargo, se espera que a lo largo del curso se pueda ir mejorando estos resultados con la aplicación de nuevas técnicas y métodos para la obtención y procesamiento de datos.

El código de la exploración de datos como el del modelo de recomendación pueden encontrarse en este repositorio de GitHub en la liga a continuación:

https://github.com/BigDataEquipo40/MNA-BigData-40/tree/main/Proyecto_Avance_1

Conclusión

Se puede llegar a la conclusión de que existen diversas formas de crear un sistema de recomendación; sin embargo, este documento plantea la base de como podemos

realizarlo y también nos muestra puntos a tomar en consideración para realizar mejoras en nuestro modelo.

Entre las mejoras que se proponen es agregar otras variables como la categoría del producto, marca e inclusive el sexo de las personas con el fin de tener una recomendación más personalizada, además de un tratamiento mayor de los datos para tratar el desbalance que existe en la cantidad de ratings que existen y que puedan provocar un mayor sesgo. Por lo que en una siguiente entrega se espera obtener mejores resultados con los métodos y estrategias que se presenten para tratar este tipo de casos.

Referencias:

MengtingWan. (n.d.). *GitHub - MengtingWan/marketBias: data and source code for "Addressing Marketing Bias in Product Recommendations" WSDM'20*. GitHub. https://github.com/MengtingWan/marketBias/tree/master

Mulyawan, B., Christanti, M. V., & Wenas, R. (2019). Recommendation Product Based on Customer Categorization with K-Means Clustering Method. *IOP Conference Series. Materials Science and Engineering*, *508*, 012123. https://doi.org/10.1088/1757-899x/508/1/012123

Wang, W., & Lu, Y. (2018). Analysis of the mean Absolute Error (MAE) and the Root Mean square Error (RMSE) in assessing rounding model. *IOP Conference Series. Materials Science and Engineering*, *324*, 012049. https://doi.org/10.1088/1757-899x/324/1/012049