PROYECTO FINAL

Sistema de recomendación de películas

Proyecto final Data Science

Alumno:
Albert Vila

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

< Índice >

Sistema de recomendación de películas

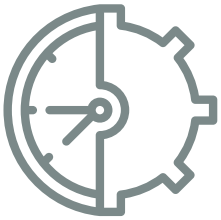
Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

< Necesidades clave >

En la era digital actual, la cantidad de información disponible en línea es abrumadora y continúa creciendo a un ritmo exponencial. Ya sea en plataformas de streaming, tiendas onlines o redes sociales. Los usuarios se encuentran ante una vasta cantidad de opciones y, por lo tanto, tomar decisiones informadas sobre qué película o serie ver, se vuelve cada vez más desafiante. Aquí es donde los sistemas de recomendación se vuelven esenciales, respondiendo a varias necesidades clave:

1. Personalización:

- Contexto del Usuario: Los sistemas de recomendación permiten que las plataformas ofrezcan experiencias personalizadas a cada usuario, adaptando su contenido según las preferencias y comportamientos individuales.
- Relevancia: Aseguran que las opciones presentadas al usuario sean relevantes y de su interés, mejorando así su experiencia de usuario y satisfacción.



2. Navegación Eficiente:

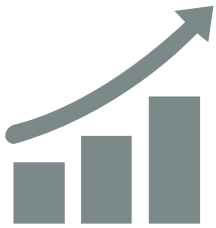
- Facilitar Decisiones: Ayudan a los usuarios a navegar a través de la amplia gama de opciones disponibles, proporcionándoles selecciones que probablemente encuentren atractivas.
- Ahorro de Tiempo: Reducen el tiempo que los usuarios necesitan para buscar y decidir sobre un producto o servicio, filtrando y destacando las opciones más pertinentes.

3. Aumento de la Retención de Usuarios:

- Engagement: Al ofrecer recomendaciones personalizadas, los usuarios son más propensos a interactuar y pasar más tiempo en la plataforma.
- Lealtad: Los sistemas de recomendación mejoran la lealtad del usuario al hacer que se sientan valorados y comprendidos en sus preferencias y necesidades.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar a los usuarios de las plataformas de streaming



4. Incremento en Ventas y Utilización:

- Descubrimiento: Facilitan el descubrimiento de nuevos contenidos que los usuarios podrían no haber encontrado por sí mismos.
- Conversión: Mejoran las tasas de conversión al dirigir a los usuarios hacia contenidos o productos que son más propensos a comprar o consumir.

5. Innovación y Competitividad:

- Innovación: Los sistemas de recomendación son una herramienta de innovación, permitiendo que las plataformas ofrezcan nuevas y mejoradas experiencias a sus usuarios.
- Ventaja Competitiva: En mercados saturados, tener un sistema de recomendación eficiente puede diferenciar a una plataforma de sus competidores, proporcionando una ventaja en términos de retención de usuarios y satisfacción del cliente.

Los sistemas de recomendación, por lo tanto, no solo enriquecen la experiencia del usuario, sino que también es una herramienta estratégica para las empresas y plataformas online, permitiéndoles maximizar su engagement, optimizar sus inventarios y mejorar estrategias de marketing. Lo que finalmente se traduce en mayores ingresos y una mejor posición en el mercado.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

<Importación de Librerías>

Este proyecto desarrolla un sistema de recomendación de películas utilizando técnicas de aprendizaje automático y filtrado colaborativo. A través del análisis y manipulación de un conjunto de datos y de la implementación de modelos de factorización matricial (como SVD). Con esto buscaba predecir las calificaciones de los usuarios hacia las películas, para hacer buenas recomendaciones que satisfagan al usuario.

1. Introducción:

El objetivo de un sistema de recomendación es sugerir productos, servicios o información relevantes a los usuarios basándose en sus preferencias y patrones. En este proyecto, se exploran películas, buscando recomendar títulos que puedan ser de interés para los usuarios **basándose en sus calificaciones previas**. Tal y como funciona el mundo del streaming, cada vez más al alza, y con nuevas plataformas que se suman a la vasta oferta ya existente. Un buen sistema de recomendación es esencial, para agilizar y potenciar la experiencia del usuario, y destacar ante la competencia.

2. Metodología:

2.1. Preprocesamiento de Datos:

Cargué cuatro conjuntos de datos: **movies_metada**, **credits**, **ratings** y **links** descargados desde **Kaggle** y que proviene de **The Movies Dataset**, una base de datos de 45000 películas y 26 millones de puntuaciones de aproximadamente 270.00 usuarios que contiene información sobre películas y calificaciones dadas por los usuarios. Llevé a cabo la limpieza y transformación de los datos, eliminando duplicados, gestionando valores nulos y convirtiendo tipos de datos para que pudieran ser utilizados.

2.2. Exploración de Datos:

Exploré las características de las películas y las calificaciones, verificando la distribución de las calificaciones, la cantidad de películas por género, y la cantidad de calificaciones por usuario, entre otros.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

2.3. Filtrado Colaborativo:
Este método emplea las calificaciones de los usuarios para predecir cómo un usuario podría calificar una película que aún no ha calificado. Usé el algoritmo SVD para la factorización de la matriz de usuario-ítem y realicé una validación cruzada para evaluar el rendimiento del modelo en términos de RMSE y MAE.

3. Preprocesamiento de datos:

3.1. Preprocesamiento de Datos:
Cargué cuatro conjuntos de datos: **movies_metada**, **credits**, **ratings** y **links** descargados desde Kaggle y que proviene de **The Movies Dataset**, una base de datos de 45000 películas y 26 millones de puntuaciones de unos 270.00 usuarios. Contiene información sobre películas y calificaciones dadas por los usuarios, y en la que llevé a cabo la limpieza, transformación de datos, eliminación de duplicados, y gestione los valores nulos, convirtiendolos en tipos de datos para que pudieran ser utilizados.

4.Resultados:

4.1. Sistema de recomendación por Score y genero.
Primero vamos a valorar las películas, usando la fórmula de IMDB

$$(\frac{v}{v+m} \times R) + (\frac{m}{v+m}$$

La fórmula **calcula una media de votación,** *C*, para todas las películas. **Establece un umbral de votos,** *m*, para determinar las películas que tienen un número de votos significativo y deberían ser consideradas para recomendación, creo la columna **Score** donde aplico la fórmula de IMDB y muestro los resultados. De esta manera, he conseguido mi primer sistema de recomendación. Un Top 10 básico, de películas ordenadas según el **Score**. Luego, de nuevo, hacemos uso del **Score**, para filtrar las películas mejor valoradas, pero en este caso por género. Los resultados son buenos. Tanto el TOP 10 general como el de género, dan resultados muy coherentes.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

4.2. Filtrado por contenido

Preprocesamiento de datos. Consta de:

- Verificación de valores no numéricos en el ID.
- Eliminación de duplicados.
- Conversión de tipos de datos.
- Manejo de datos en formato de cadena para convertirlos en listas (utilizando literal_eval).
- Limpieza de datos (eliminación de espacios, extracción del nombre del director y cast, etc).

Creación de **Soup**: Se genera una "sopa" en forma de cadena de texto que contiene información del casting, director, géneros y descripción.

Vectorización y Similitud de Coseno:

- Se usa **TfidfVectorizer** para convertir la "sopa" en una matriz de características.
- Se calcula la similitud del coseno entre las películas empleando esas características.

Generación de Recomendaciones:

- Se define una función (**get_recommendations**) que toma un título de la película y devuelve películas similares basándose en la similitud del coseno.
- También se propone una función mejorada (**improved_recommendations**) que toma en cuenta la puntuación de la película y géneros para hacer recomendaciones más precisas.

44.2.1. Análisis de los Resultados:

Las recomendaciones iniciales para "The Ring" son buenas, pero también incluye películas como "Lilo & Stitch" y "Kramer vs. Kramer", que no parecen ser relevantes para un fan del género de terror. Esta es una limitación del modelo basado solo en similitud de contenido sin considerar otras métricas.

Luego, se intentan mejorar las recomendaciones considerando también el género y las puntuaciones de las películas.

La función **improved_recommendations** se define para este propósito y ofrece recomendaciones bastante más acertadas. Incluyendo películas no directamente relacionadas con el terror como "Donnie Darko" y "Zodiac", pero que sí tiene una relación tangencial con el género. Hay alguna sorpresa, como "The Bourne Identity", que es un thriller y podría no alinearse con los gustos de alguien a quien le gustan películas de terror sobrenatural, como “The Ring". Aun así, el thriller es más cercano al terror que lo que sería "Lilo & Stitch" o "Kramer vs. Kramer".
En conclusión, no observo ningún resultado que parezca fuera de lugar.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

4.3. Filtrado colaborativo

- Cargo el dataset ratings y muestro las primeras tres filas. Utilizo la biblioteca **Surprise** para implementar un filtro colaborativo usando **SVD** (Descomposición de Valores Singulares).
- Configuro el **reader** y el conjunto de datos para Surprise y escojo una muestra del 60% .
- Aplico SVD con unos parámetros específicos y evalúo el algoritmo usando la **validación cruzada** (CV = 3), y mostrando las métricas **RMSE** y **MAE.**
- Finalmente, muestro los ratings de un usuario específico (userId = 1) y realizo una predicción puntual para un usuario y una película.

4.3.1.Evaluación de Resultados:

Los valores RMSE y MAE en los resultados son relativamente bajos, lo cual es bueno. Ahora se ha de implementar en el filtro híbrido para evaluar su funcionamiento.

4.4. Filtro Híbrido

- Cargo el dataset **links**, y defino la función **hybrid**, que combina las recomendaciones basadas en contenido con las estimaciones de puntuación del filtro colaborativo.
- El filtro híbrido funciona al tomar la puntuación de similitud del filtro de contenido y estima la puntuación del usuario utilizando el filtro colaborativo. Luego devuelve una lista de recomendaciones ordenadas según la puntuación.
- Se define una función **convert_int** que trata de convertir su entrada en un entero y devuelve NaN si no es posible. Esta función se aplica a la columna **tmdbId** del dataset.
- La función hybrid toma **userId** y **title** como entradas y devuelve recomendaciones. Usa similitud coseno para encontrar películas similares a la indicada por title, luego usa el algoritmo SVD para predecir la puntuación que el usuario daría y ordena las recomendaciones según en esa puntuación.

4.4.1. Evaluación de los resultados:

En este caso, he elegido la película de ciencia ficción **Gatacca**, para poner a prueba el filtro. Al cambiar de usuario, no hay mucha diferencia en los resultados. Esto podría indicar poca variación en los gustos, y que todos los usuarios se decantan por elecciones similares o bien que hay margen de mejora, ya que pocas películas en la lista tienen una afinidad real con Gattaca.
Mirando el listado, me quedaría con **The Butterfly Effect y Cube** como mejores resultados. En cuanto al resto hay alguna coincidencia a nivel de actores y director, pero tanto los géneros como el contenido de esas películas, no son muy coincidentes con Gatacca.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

5. Conclusiones

- **Validación de los Resultados:** Los sistemas de recomendación implementados, tanto los de género como de contenido, han demostrado ser técnicamente viables, generando recomendaciones lógicas y predicciones en las calificaciones con un error moderado. Sin embargo, las recomendaciones y las predicciones deben validarse adicionalmente en un contexto práctico para asegurar que sean percibidas como relevantes y precisas por los usuarios finales. En el sistema híbrido, por otro lado, las recomendaciones no tienen tanta lógica y hay mucha similitud entre los diferentes usuarios.

- **Mejoras Futuras:** Aunque los modelos presentados ofrecen una base sólida, hay varias rutas para la mejora y la optimización. Esto podría incluir la exploración de otros algoritmos de recomendación para mejorar los resultados que he obtenido con el filtro híbrido, o la inclusión de más características en los modelos basados en contenido.

- **Implicaciones Prácticas:** En términos prácticos, el sistema podría implementarse en una plataforma de streaming para ayudar a los usuarios a descubrir nuevas películas basadas en sus preferencias pasadas y las características de las películas. Sin embargo, la implementación práctica también debe considerar factores como la diversidad, la novedad y la popularidad de las recomendaciones para asegurar que los usuarios estén constantemente expuestos a nuevas opciones.

En conclusión. El proyecto ha demostrado cómo implementar un sistema de recomendación utilizando técnicas de aprendizaje automático. El modelo basado en el género y el de contenido proporcionan recomendaciones bastante buenas y coherentes, mientras que el modelo de filtrado colaborativo, aunque es capaz de predecir las calificaciones de los usuarios. Al implementarlo en el filtro híbrido, aunque se cambie de usuario, genera resultados demasiado similares, lo cual plantea dudas.
El RMSE y MAE obtenidos en el modelo de filtrado colaborativo indican que, aunque el modelo se mueve dentro de unos buenos parámetros, esto no corresponde con el resultado obtenido. He ajustado los hiperparámetros usando **Grid Search** para optimizar el modelo, pero solo he obtenido una ligera mejora. Quizás emplear un conjunto de datos más grande, seguir variando los hiperparámetros, o bien utilizar algoritmos diferentes, serían soluciones que podrían ayudar a mejorar las predicciones.

Sistema de recomendación de películas

Un proceso más eficiente y personalizado para fidelizar los usuarios de las plataformas de streaming

6. Referencias:

- https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
- https://www.kaggle.com/code/ibtesama/getting-started-with-a-movie-recommendation-system
- https://www.aprendemachinelearning.com/sistemas-de-recomendacion/
- https://hescaso.github.io/recomendador/