Analítica III - Caso de Estudio Marketing

Cesar Iván Ávila Díaz. cesar.avila@udea.edu.co
Verónica Andrea Morales González. veronica.moralesg@udea.edu.co

Repositorio GitHub: https://github.com/veronica1908/MARKETING_E2_A_III

Resumen: Se presenta una solución diseñada para mejorar la experiencia de los usuarios de una plataforma en línea a través de un sistema de recomendación de películas. Se inicia con la extracción, limpieza y transformación de datos de la base de datos bd_movies, seguido de la aplicación de filtros para mejorar la calidad de las recomendaciones. Se desarrollan diferentes sistemas de recomendación, incluyendo aquellos basados en popularidad, contenido y filtrado colaborativo, adaptados a distintos tipos de usuarios. Finalmente, se implementa un sistema de calificación de películas para enriquecer las recomendaciones y se evalúan varios modelos para garantizar su eficacia, los cuales son almacenados de manera local en extensiones de excel y csv.

Palabras Clave: Calificaciones, Películas, Plataforma online, Sistema de recomendación, Usuarios.

1. Descripción del caso problema

Una plataforma online quiere tener una solución que le permita hacer recomendaciones de películas a sus usuarios, con el objetivo de que estos tengan una mejor experiencia, y esto permita mejorar su fidelización y recomendación a nuevos clientes. Cada equipo de trabajo será el encargado de diseñar un sistema que permita hacer recomendaciones a los clientes.

2. Diseño de la solución para el caso de estudio

a. Extraer la información:

Este paso implica obtener los datos necesarios para el sistema de recomendación y el sistema de calificación de películas. Puede incluir la recopilación de datos de películas, información de usuarios, historiales de visualización y calificaciones, entre otros.

b. Realizar limpieza y transformación de los datos:

En este paso, se lleva a cabo la limpieza y transformación de los datos obtenidos. Esto puede incluir la eliminación de datos irrelevantes o duplicados, el tratamiento de valores nulos o faltantes, y la conversión de datos en un formato adecuado para su análisis y procesamiento.

c. Aplicar filtros para mejorar la confiabilidad y calidad de las recomendaciones:

Se aplican filtros y técnicas de preprocesamiento de datos para mejorar la calidad y confiabilidad de las recomendaciones. Esto incluye la creación de tablas que reúnen solo los datos con la cantidad requerida de información relevante para el análisis.

d. Diseñar Sistemas de Recomendación:

En este paso, se diseñan e implementan los sistemas de recomendación basados en los requisitos y las necesidades del usuario. Esto puede incluir sistemas basados en contenido, filtrado colaborativo, y técnicas avanzadas como aprendizaje profundo o sistemas híbridos.

e. Determinar el Tipo de Usuario y Aplicar el Sistema de Recomendación Correspondiente:

Dependiendo del tipo de usuario (principiante, intermedio o avanzado), se aplica el sistema de recomendación adecuado:

Principiante: <= 50 películas vistas - Aplicar sistema de recomendación basado en popularidad.

Intermedio: más de 50 películas solo vistas - Aplicar sistema de recomendación basado en todo lo visto por el usuario.

Avanzado: más de 50 películas vistas y calificadas - Aplicar sistema de recomendación basado en todo lo visto por el usuario y sus calificaciones.

f. Implementar Sistema de Calificación de las Películas:

Se implementa un sistema que permite a los usuarios calificar las películas después de verlas. Esto puede incluir la creación de una interfaz de usuario para las calificaciones y el almacenamiento de las calificaciones proporcionadas por los usuarios para su posterior uso en el sistema de recomendación.

Estos pasos proporcionan una guía estructurada para el diseño e implementación de un sistema de

recomendación y un sistema de calificación de películas, con el objetivo de mejorar la experiencia del usuario y aumentar la satisfacción y fidelización de los clientes.

Se muestra gráficamente la propuesta de solución

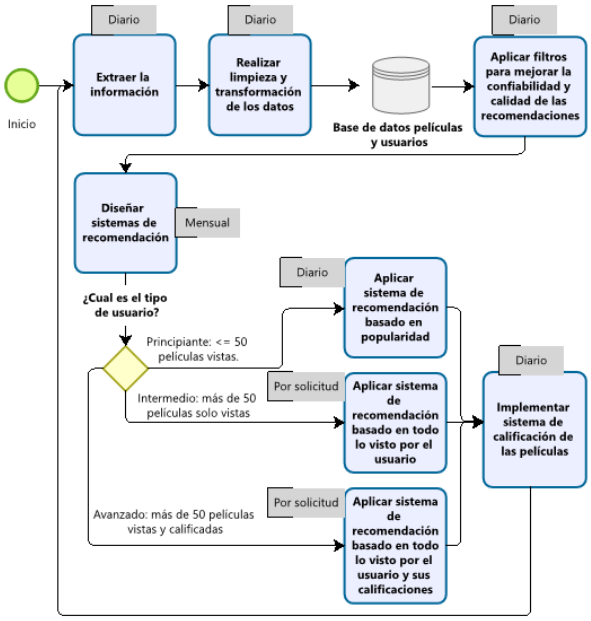


Ilustración 1. Solución teórica propuesta. Elaboración propia.

3. Desarrollo del caso de estudio

a. Extraer la información:

En este paso, se recopilan datos de películas, usuarios y sus interacciones con las películas desde la base de datos bd_movies.

b. Realizar limpieza y transformación de los datos:

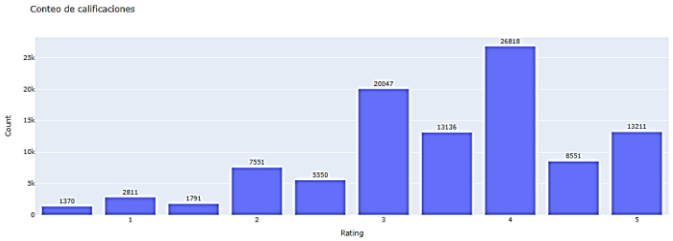
Durante esta etapa, se realizan operaciones de limpieza y transformación de datos para asegurar la coherencia y calidad de los datos. Esto incluye manejar valores nulos, eliminar duplicados y ajustar el formato de los datos según sea necesario. Por ejemplo, en la tabla movies se realizó la limpieza y transformación de datos al dividir los géneros de las películas en listas separadas para posteriormente generar tabla de dummies que aporten al sistema de recomendación, se extrajo también el año de publicación desde el nombre de la película.

c. Aplicar filtros para mejorar la confiabilidad y calidad de las recomendaciones:

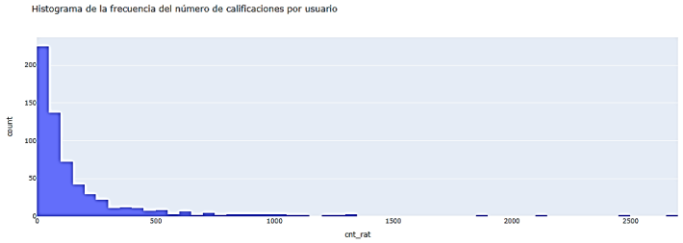
En esta fase, se aplicaron filtros para mejorar la calidad de las recomendaciones: se creó una tabla con usuarios con más de 50 películas vistas (usuarios_sel), una tabla con películas que han sido calificadas por más de 50 usuarios (movies_sel), tablas filtradas de calificaciones con usuarios y películas (ratings_final), tabla con esta información seleccionada anteriormente (movies_final) y finalmente la tabla full_ratings que contiene la información cruzada de las tablas anteriores con las siguientes variables: user_id, movie_id, rating, movie_title, genres y year.

Además se verifica el comportamiento de los datos a través de la exploración de los mismos, para lo cual se obtiene lo siguiente:

En la tabla ratings se analiza la distribución por calificación (rating); el rating 4.0 es el que tiene mayor cantidad de calificaciones con 26.818 y el que menos calificaciones tiene es el rating 0.5 con 1370 calificaciones.



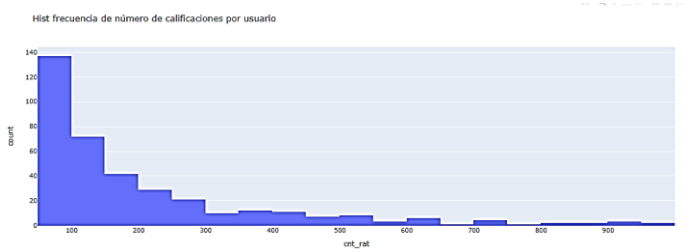
En cuanto a cantidad de calificaciones por usuario, el usuario 414 es el que ha realizado mayor cantidad de calificaciones con un total de 2698 películas. El usuario 53 es el que menos películas ha calificado, con un total de 20. La cantidad promedio de calificaciones por usuario es de aproximadamente 165 películas. El 25% de los usuarios han calificado 35 películas o menos. El 50% de los usuarios han calificado 70 películas o menos.



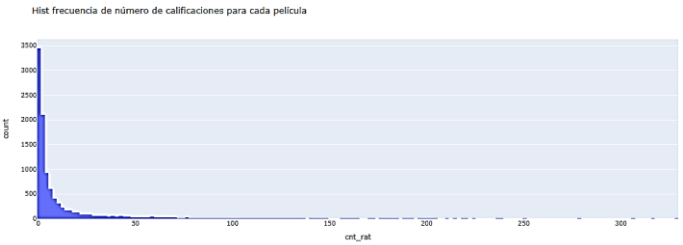
Al aplicar un filtro para usuarios que tengan más de 50 calificaciones se obtiene que hay un total de 373 usuarios que tienen entre 50 y 1000 calificaciones. El promedio de calificaciones por usuario es aproximadamente 201.86, lo que sugiere que, en promedio, los usuarios han calificado alrededor de 201 películas.

El usuario que ha aportado menor número de calificaciones, ha aportado 50 calificaciones, mientras que el usuario que ha aportado el mayor número de calificaciones, ha aportado 977. Esto muestra una amplia variabilidad en el número de calificaciones entre los usuarios.

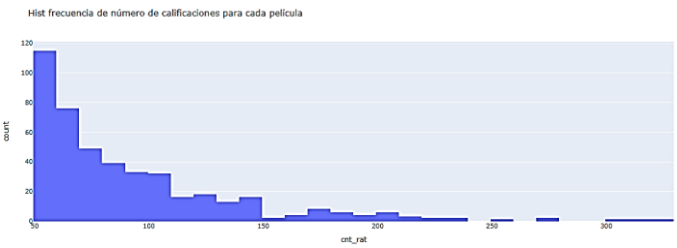
En general, la mayoría de los usuarios han aportado una cantidad considerable de calificaciones, con una variabilidad significativa en la cantidad de calificaciones entre los usuarios. En el histograma, se evidencia que hay una mayor cantidad de usuarios que han calificado entre 50 y 99 películas; 137 en total. Le siguen 72 usuarios que han calificado entre 100 y 149 películas.



En cuanto a cantidad de calificaciones por película se tiene que la cantidad promedio de calificaciones por película es de aproximadamente 10. El número mínimo de calificaciones por película es 1, lo que indica que hay usuarios que han calificado muy pocas películas, mientras que el número máximo de calificaciones por película es de 9724, lo que indica que hay usuarios que han calificado un gran número de películas y que el rango es demasiado alto.



Igualmente se aplicó el filtro de solo películas cuya cantidad de calificaciones sea igual o mayor que 50 y se obtuvo que son 450 películas las que cumplen y su promedio de calificaciones es de casi 92. El número mínimo de calificaciones por el filtro es de 50 y el máximo es de 329. De acuerdo con el histograma, se tiene que: hay 115 películas que cuentan con un número de calificaciones entre 50 y 59. Seguidas de 76 películas que cuentan con un número de calificaciones que va de 60 a 69. Solo hay una película que tiene un número de calificaciones entre 320 y 329.



Además se analiza el género con mejor promedio de calificación y el género de película con mejor promedio de calificación es Action|Adventure|Comedy|Fantasy|Romance con 4.24, mientras que el de peor promedio de calificación es Action|Comedy|Sci-Fi|Western con 2,13.

d. Diseñar Sistemas de Recomendación:

Aquí es donde se desarrollan los sistemas de recomendación, en una primera parte se exploran recomendaciones basados en popularidad:

	movie_title	avg_rat	w_num
0	Shawshank Redemption, The (1994)	4.421739	230
1	Godfather, The (1972)	4.298780	164
2	Dr. Strangelove or: How I Learned to Stop Worr...	4.266304	92
3	Cool Hand Luke (1967)	4.259615	52
4	Schindler's List (1993)	4.253012	166
5	Chinatown (1974)	4.250000	54
6	Princess Bride, The (1987)	4.246032	126
7	Casablanca (1942)	4.244444	90
8	Goodfellas (1990)	4.243478	115
9	Apocalypse Now (1979)	4.240196	102

La película "Shawshank Redemption, The (1994)" tiene el mejor promedio de calificación (4,44) y un total de 219 calificaciones recibidas.

Las películas más vistas con promedio de las calificaciones y cantidad de calificaciones

	movie_title	avg_rat	w_num
0	Forrest Gump (1994)	4.121622	259
1	Pulp Fiction (1994)	4.209544	241
2	Shawshank Redemption, The (1994)	4.421739	230
3	Matrix, The (1999)	4.132159	227
4	Silence of the Lambs, The (1991)	4.158257	218
5	Star Wars: Episode IV - A New Hope (1977)	4.206731	208
6	Jurassic Park (1993)	3.774112	197
7	Star Wars: Episode V - The Empire Strikes Back...	4.225543	184
8	Fight Club (1999)	4.234973	183
9	Terminator 2: Judgment Day (1991)	4.000000	182

Forrest Gump (1994) es la película más vista, o al menos la más veces calificada, sin embargo su promedio de calificación no estuvo en el top 10 de la consulta anterior, su promedio es de 4.12 con 247 calificaciones.

Películas con mejor promedio de calificación y año de publicación

	year	movie_title	avg_rating
0	2016	Deadpool (2016)	3.755556
1	2014	Guardians of the Galaxy (2014)	4.039216
2	2014	Interstellar (2014)	3.885965
3	2014	Grand Budapest Hotel, The (2014)	3.775510
4	2013	Wolf of Wall Street, The (2013)	3.847826
..	...	...	...
431	1941	Citizen Kane (1941)	4.030303
432	1940	Fantasia (1940)	3.770833
433	1940	Pinocchio (1940)	3.410714
434	1939	Wizard of Oz, The (1939)	3.886364
435	1937	Snow White and the Seven Dwarfs (1937)	3.605634

Al ver la información por año de lanzamiento, se tiene que la película más recientemente lanzada es Deadpool en 2016, sin embargo, su promedio de

calificación no está entre los más altos, siendo de 3.71. Al observar la película con año de lanzamiento más antiguo, se observa que tiene incluso menor promedio de calificación en comparación con la más reciente: Snow White and the Seven Dwarfs (1937) con 3.60.

Se explora también Sistema de recomendación basado en contenido un solo producto – Manual

Para ello, se utiliza la tabla de dummies que contiene el año escalado y el género separado, se aplica un ejemplo con la película 'Good Time (2017)' para ver las 15 recomendaciones de películas similares y se obtienen las siguientes:

Las 15 películas recomendadas son:	correlación	
9577	1.000000	Good Time (2017)
9465	1.000000	T2: Trainspotting (2017)
9618	1.000000	Three Billboards Outside Ebbing, Missouri (2017)
9714	1.000000	Dogman (2018)
9352	1.000000	The Infiltrator (2016)
9357	1.000000	Hell or High Water (2016)
8952	1.000000	Wild Horses (2015)
8928	1.000000	Black Mass (2015)
8991	1.000000	Irrational Man (2015)
8311	0.999999	American Hustle (2013)
8188	0.999999	Bling Ring, The (2013)
7967	0.999999	Lawless (2012)
8129	0.999999	Place Beyond the Pines, The (2012)
7227	0.999998	Staten Island (2009)
7226	0.999998	Prophet, A (Un Prophète) (2009)

Se aplica también un Sistema de recomendación basado en contenido KNN un solo producto visto

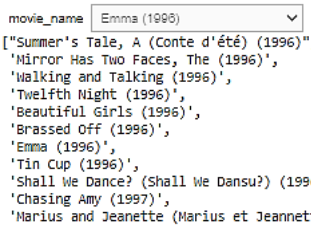
Se aplica el ejemplo para la película 'Spider-Man (2002)'

Este modelo utiliza la distancia coseno para medir la similitud entre los puntos en un espacio de características y encuentra los vecinos más cercanos para cada punto en función de esta medida de similitud. Esto es útil para identificar películas similares en función de sus características, representadas en la tabla de dummies.

Se obtiene las siguientes 10 recomendaciones:

```
['Clockstoppers (2002)',
 'Spider-Man (2002)',
 'X2: X-Men United (2003)',
 'Jurassic Park III (2001)',
 'I, Robot (2004)',
 'War of the Worlds (2005)',
 'Stealth (2005)',
 'Lost World: Jurassic Park, The (1997)',
 'Spawn (1997)',
 'Babylon A.D. (2008)',
 'Death Race (2008)']
```

De esta manera se habilita una pestaña interactiva para seleccionar una película y que se den las 10 recomendaciones con base en la selección.



Para la segunda parte, se considera el sistema de recomendación basado en contenido KNN con base en lo visto por el usuario, para el ejemplo, se selecciona al usuario 1, se filtran los ratings del usuario seleccionado, las películas vistas y las películas no vistas por este usuario, se entrena el modelo y se obtienen las recomendaciones siguientes:

	title	movieId
9698	Love, Simon (2018)	184997
9718	Boundaries (2018)	189043
9446	A Dog's Purpose (2017)	167380
9650	Daddy's Home 2 (2017)	180231
9653	The Disaster Artist (2017)	180297
9524	Rough Night (2017)	171867
9492	Table 19 (2017)	170401
9614	Little Boxes (2017)	176805
9305	The Meddler (2016)	159077
9451	A Street Cat Named Bob (2016)	167732
9249	Florence Foster Jenkins (2016)	155659

Además se agrega de manera interactiva la selección del usuario para que traiga las recomendaciones correspondientes:

Se implementa también el sistema de recomendación por filtro colaborativo, se define varios modelos de filtrado colaborativo basado en vecinos más cercanos (KNN), cada uno con sus propios ajustes, se utiliza validación cruzada para evaluar cada modelo en términos de error medio absoluto (MAE) y error cuadrático medio (RMSE). Esto se hace para cada modelo en el conjunto de modelos definidos y se obtiene lo siguiente:

	MAE	RMSE	fit_time	test_time
knns.KNNBaseline	0.626257	0.822691	0.101353	1.467151
knns.KNNWithZScore	0.631205	0.831617	0.177952	1.779076
knns.KNNWithMeans	0.635309	0.833818	0.091351	1.219660
knns.KNNBasic	0.664325	0.874157	0.064550	1.136549

MAE (Error Medio Absoluto): Esta métrica representa la magnitud promedio del error entre las calificaciones reales y las predicciones del modelo. Cuanto menor sea el valor del MAE, mejor será el rendimiento del modelo en términos de precisión de las predicciones. En este caso, el modelo KNNBaseline tiene el MAE más bajo (0.626), seguido por KNNWithZScore

(0.631), KNNWithMeans (0.635) y KNNBasic (0.664).

RMSE (Error Cuadrático Medio): Esta métrica mide la raíz cuadrada de la diferencia cuadrática promedio entre las calificaciones reales y las predicciones del modelo. Al igual que con el MAE, cuanto menor sea el valor de RMSE, mejor será el rendimiento del modelo. En este caso, los resultados son similares, donde el modelo KNNBaseline tiene el RMSE más bajo (0.822), seguido por KNNWithZScore (0.832), KNNWithMeans (0.834) y KNNBasic (0.874).

En conclusión, el modelo KNNBaseline parece ser el mejor en términos de precisión de predicción, seguido por KNNWithZScore, KNNWithMeans y KNNBasic. Sin embargo, la elección del mejor modelo también puede depender de otros factores como el tiempo de ajuste (fit_time) y el tiempo de evaluación (test_time), que deben considerarse junto con las métricas de rendimiento. En este caso, el modelo KNNBaseline tiene un tiempo de ajuste más rápido, pero un tiempo de evaluación más largo en comparación con otros modelos.

Finalmente se crea la función para recomendar las 10 películas con mejores predicciones y llevar a la base de datos para consultar el resto de información.

e. Determinar el tipo de usuario y aplicar el sistema de recomendación correspondiente:

Basado en el análisis de datos de los usuarios (por ejemplo, la cantidad de películas vistas y calificadas), se clasificarían en principiantes, intermedios o avanzados. Luego, se aplicarían los sistemas de recomendación adecuados a cada tipo de usuario, utilizando la información recopilada de bd_movies y los modelos desarrollados en el paso anterior.

Aquí, para el despliegue se tiene en cuenta un ejecutable que reúne las bases y los modelos diseñados, lo cual arroja las recomendaciones correspondientes que son almacenadas de manera local en un documento excel y un documento .csv con una lista de usuarios que fue limitada para cuatro en este caso.

f. Implementar Sistema de Calificación de las Películas:

Se considera una funcionalidad para que los usuarios califiquen las películas después de verlas. Estas calificaciones se integran en el análisis de datos para mejorar las recomendaciones futuras.

4. Conclusiones y recomendaciones

Se puede concluir que la implementación de un sistema de recomendación de películas es factible y puede mejorar significativamente la experiencia de los usuarios en la plataforma online. Se identificaron diversas técnicas y modelos, como el filtrado colaborativo y el sistema basado en contenido, que pueden adaptarse según el nivel de experiencia del usuario. Además, se recomienda continuar mejorando el sistema mediante la recopilación continua de datos de calificaciones de usuarios y la optimización de los algoritmos de recomendación para ofrecer sugerencias más precisas y personalizadas.

Desde el punto de vista de marketing, se pueden ofrecer las siguientes recomendaciones estratégicas:

- ✓ Campañas de promoción y recomendación: Implementar campañas de marketing que destaquen las características únicas del sistema de recomendación de películas, resaltando su capacidad para ofrecer sugerencias relevantes y adaptadas a los intereses individuales de cada usuario.
- ✓ Incentivos para la calificación de películas: Ofrecer incentivos a los usuarios para que califiquen las películas después de verlas, como descuentos en futuras suscripciones o acceso exclusivo a contenido adicional. Esto puede ayudar a aumentar la cantidad y calidad de las calificaciones, lo que a su vez mejora la precisión del sistema de recomendación.
- ✓ Colaboraciones con influencers y críticos de cine: Establecer asociaciones con influencers y críticos de cine reconocidos para promocionar el sistema de recomendación de películas y generar interés entre nuevos usuarios. Las reseñas y recomendaciones de personas influyentes pueden tener un gran impacto en la percepción y la adopción del sistema.
- ✓ Análisis continuo de datos y retroalimentación del usuario: Realizar un seguimiento constante del rendimiento del sistema de recomendación y recopilar comentarios de los usuarios para identificar áreas de mejora. Esto permite ajustar y optimizar continuamente el sistema para satisfacer las necesidades y expectativas cambiantes de los usuarios.