INFORME CASO DE ESTUDIO MARKETING

Diseño de la solución

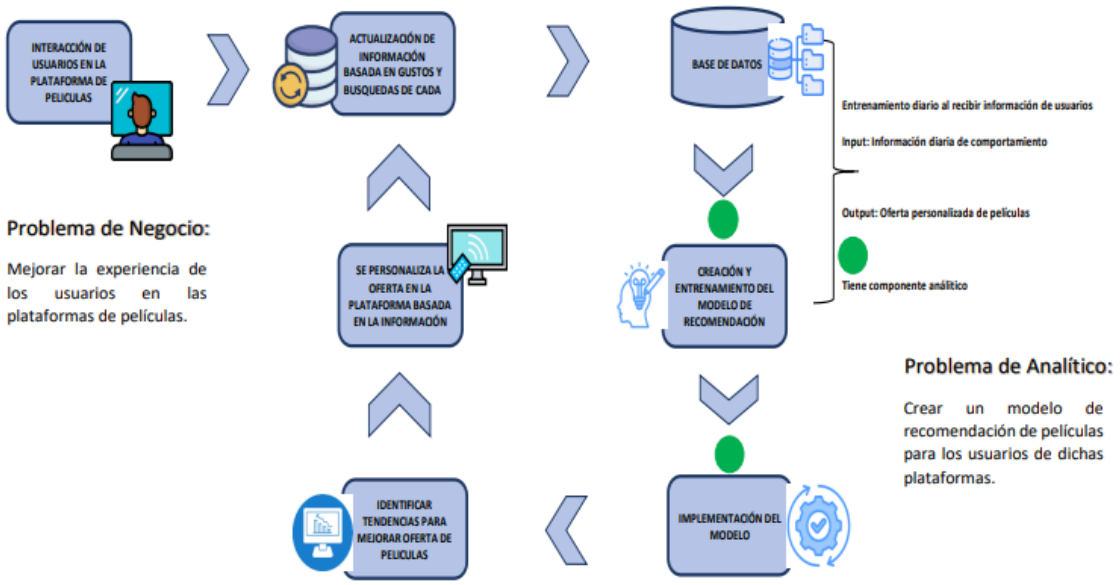


Imagen 1. Diseño de la solución

Para el desarrollo y diseño de la solución se tuvieron en cuenta los diferentes problemas a intervenir, siendo el reto como negocio mejorar la experiencia de los usuarios de plataformas de películas, así mismo su componente analítica pretende crear un modelo de recomendación personalizado de películas. Para la creación y entrenamiento de dicho modelo se hizo una limpieza y transformación de datos y se entrenó el modelo con la base de datos resultante, se hizo un estudio del comportamiento de los usuarios para así determinar gustos y comportamientos. Posteriormente se seleccionaron las variables que serían representativas para alimentar los diferentes sistemas, siendo los anteriores basados en la popularidad del filme, contenido KNN y KNN visto todo por un usuario y finalmente un sistema de recomendación basado en un filtrado colaborativo. Una vez hechas las recomendaciones se actualiza la oferta para el usuario y se hacen las nuevas recomendaciones basadas en su comportamiento y preferencias.

Limpieza y transformación

En la limpieza de las bases de datos se realizó un primer filtro a la base de ratings en donde solo se dejaron las películas que han sido calificadas por más de 10 usuarios, debido a que las recomendaciones deben de estar basadas por un buen número de calificaciones por los usuarios, además se crea una tabla ‘movies_f’ en la cual solo quedan las películas que cumplen la condición de que hayan sido calificadas más de 10 veces por los usuarios. Finalmente se crea una tabla en donde se une el rating con la información respectiva a cada película.

Análisis exploratorio

Al analizar las dos tablas movies y ratings no se encuentran valores duplicados y nulos, además observó el número de variables y sus características. Se realizaron las siguientes gráficas para observar el comportamiento de los usuarios en la plataforma de películas.

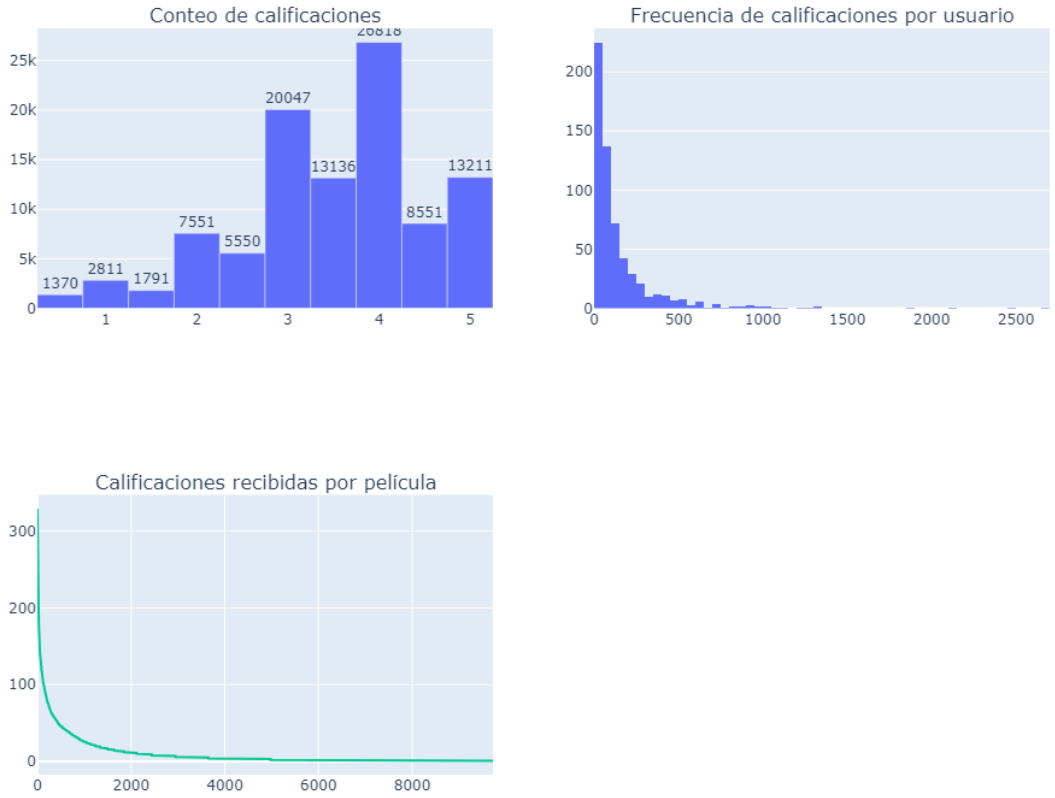


Imagen 2. Comportamiento de los usuarios

El gráfico de barras ilustra la distribución de las calificaciones, mostrando que las calificaciones de 4 y 3 tienen la mayor frecuencia, mientras que las calificaciones de 0.5 y 1.5 son las menos comunes.

En el histograma de frecuencia de calificaciones se evidencia la distribución de la cantidad de calificaciones otorgadas por los usuarios. Destaca que el usuario que más califica tiene un número de calificaciones de 2698 películas, mientras que la cantidad mínima de calificaciones otorgadas por un usuario es de 20.

Se puede interpretar del gráfico de calificaciones recibidas por película en que hay una variación considerable en el número de calificaciones asignadas a cada película. Por ejemplo, la película más popular ha recibido 329 calificaciones, mientras que hay películas que solo han sido calificadas una vez, por esta razón se decide filtrar las películas que no tengan más de 10 calificaciones recibidas

Selección de variables

En cuanto a la selección de las variables, puesto que no había un gran número de estas, simplemente se eliminó una variable que no otorgaba información relevante, puesto que se repetía en la tabla. Esta variable eliminada es: “movieId:1”. Agregando que la variable de género de las películas se separó para que cada género trabajara como una variable independiente.

Selección de sistemas de recomendación

Se seleccionaron 4 sistemas de recomendaciones, entre estos están el de recomendaciones en base a la popularidad, basado en contenido, basado en todo lo visto por el usuario y por último el sistema de recomendación de filtro colaborativo.

Afinamiento de hiperparámetros

Como se puede observar en la imagen 3 a continuación los modelos knns.KNNWithMeans y knns.KNNWithZScore para recomendaciones de filtro colaborativo tienen el RMSE muy similar, siendo este último el de menor RMSE. Sin embargo, se decide escoger el knns.KNNWithMeans, puesto que no se considera que haya una diferencia considerable entre estos dos modelos, y en cambio sí se tiene un procesamiento más rápido, tal y como se puede observar en la variable fit_time y test_time, haciendo el modelo mucho más eficiente.

	MAE	RMSE	fit_time	test_time
knns.KNNBasic	0.694122	0.904114	0.459042	1.900243
knns.KNNWithMeans	0.652995	0.853711	0.362282	2.058420
knns.KNNWithZScore	0.648666	0.852651	0.449781	2.338434
knns.KNNBaseline	0.641490	0.839511	0.505417	2.548960

Imagen 3. Métricas de los modelos

Análisis de recomendaciones

- Sistema basado en popularidad:

Estas recomendaciones son las más sencillas de realizar al solo ser analitica descriptiva, sin embargo son muy útiles y muy usadas por plataformas como spotify en sus tops musicales. En este caso creamos 3 tablas con recomendaciones, la primera las mejor calificadas con más de 100 calificaciones, la segunda es el top de películas con mayor número de calificaciones y por último una lista de la película mejor calificada por cada año con más de 50 calificaciones.

- Sistema basado en contenido KNN para una sola película

Para este modelo se busca recomendar las películas que tengan una alta correlación con otra película en particular, en general que sea de contenido similar. Este modelo es entrenado en base a las variables de ‘género’ y ‘año de estreno’, sin embargo para mejorar

más estas correlaciones se buscaría entrenar el modelo con más atributos que puedan ser representativos de las películas como el director, actores, entre otros. A continuación se puede ver las recomendaciones que el modelo da para la película de 'Toy Story':

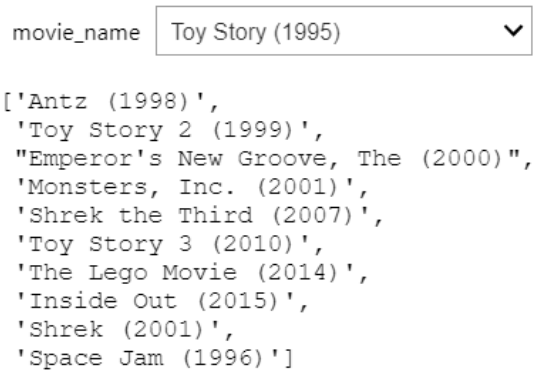
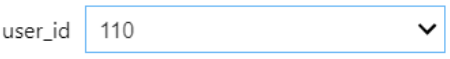


Imagen 4. Recomendaciones contenido KNN un solo producto

- Sistema basado en contenido KNN con todo lo visto por un usuario

En este modelo se busca crear recomendaciones personalizadas para cada usuario en base a su historial de películas vistas en la plataforma mediante un centroide que determinará qué películas pueden ser del agrado de cada usuario. A continuación se muestran las películas recomendadas para el 'user_id' número 110 según el modelo:



	title	movieId
62	Up Close and Personal (1996)	140
16	Money Train (1995)	20
1766	28 Weeks Later (2007)	53000
1531	Nausicaä of the Valley of the Wind (Kaze no ta...	7099
1413	xXx (2002)	5507
1777	Harry Potter and the Order of the Phoenix (2007)	54001
1673	Kiss Kiss Bang Bang (2005)	38061
1644	Kingdom of Heaven (2005)	33162
1627	Million Dollar Baby (2004)	30707
698	French Connection, The (1971)	1953
1550	Starsky & Hutch (2004)	7325

Imagen 5. Recomendaciones contenido KNN todo lo visto por el usuario

- Sistema de recomendación filtro colaborativo

En este modelo, se utilizan las calificaciones de los usuarios para sugerir películas al usuario seleccionado, priorizando las estimaciones más altas. Este enfoque es más fiable cuando hay un mayor número de calificaciones por película.

En la evaluación de este modelo para realizar predicciones, se consideran métricas como MAE (Error Absoluto Medio), RMSE (Error Cuadrático Medio), tiempo de ajuste (fit time) y tiempo de prueba (test time) y como ya se había mencionado se escoge el modelo knns.KNNWithMeans el cual presenta mejor rendimiento en los ajustes de tiempo y tiempo de prueba. En la siguiente imagen se muestran las recomendaciones dadas para el 'user_id' número 500:

	index	iid	est	title
0	995313	1221	4.191539	Godfather: Part II, The (1974)
1	994746	904	4.181347	Rear Window (1954)
2	994993	750	4.165421	Dr. Strangelove or: How I Learned to Stop Worr...
3	994823	4226	4.149265	Memento (2000)
4	994537	1089	4.132204	Reservoir Dogs (1992)
5	995867	741	4.129814	Ghost in the Shell (Kôkaku kidôtai) (1995)
6	994492	47	4.122168	Seven (a.k.a. Se7en) (1995)
7	994783	2019	4.112080	Seven Samurai (Shichinin no samurai) (1954)
8	994572	1617	4.104811	L.A. Confidential (1997)

Imagen 6. Recomendaciones modelo filtro colaborativo

Despliegue del modelo

El despliegue se realiza con temporalidad diaria, recibiendo información sobre el comportamiento de cada usuario, se realiza un entrenamiento diario. Además, en colaboración con desarrolladores, se decide implementar un diseño en la página de inicio de la plataforma web que incluya una sección dedicada a las recomendaciones. De esta manera, todas los días se recopila la información sobre las películas calificadas por los usuarios para entrenar el modelo y se genera una recomendación para todos los usuarios. Además se puede hacer una sección más general donde se les recomienda a todos cuales son las películas mejor calificadas del año que puedan acceder a ella fácilmente. Esta recomendación se publica en una base de datos específica definida por los desarrolladores web, lo que determina la información que verá el usuario en la plataforma de streaming. En la siguiente imagen se muestra como podría ser esa base para los desarrolladores puedan publicar las recomendaciones personalizadas, mostrando por efectos prácticos las 10 películas recomendadas para los 'user_id' números 1, 2 y 3:

	movieId	title	user_id
0	45517	Cars (2006)	1
1	72011	Up in the Air (2009)	1
2	4121	Innerspace (1987)	1
3	3259	Far and Away (1992)	1
4	67734	Adventureland (2009)	1
5	68554	Angels & Demons (2009)	1
6	1251	8 1/2 (8½) (1963)	1
7	5992	Hours, The (2002)	1
8	4321	City Slickers (1991)	1
9	4011	Snatch (2000)	1
10	27808	Spanglish (2004)	1
0	58047	Definitely, Maybe (2008)	2
1	45672	Click (2006)	2
2	40815	Harry Potter and the Goblet of Fire (2005)	2
3	36529	Lord of War (2005)	2
4	4886	Monsters, Inc. (2001)	2
5	3271	Of Mice and Men (1992)	2
6	2867	Fright Night (1985)	2
7	2951	Fistful of Dollars, A (Per un pugno di dollari) (1964)	2
8	1566	Hercules (1997)	2
9	1587	Conan the Barbarian (1982)	2
10	991	Michael Collins (1996)	2
0	111362	X-Men: Days of Future Past (2014)	3
1	69122	Hangover, The (2009)	3
2	4367	Lara Croft: Tomb Raider (2001)	3
3	86882	Midnight in Paris (2011)	3

Imagen 7. Recomendaciones del despliegue