Modelo recomendador de películas basado en recomendaciones de IMBD

Paolo Grimaldi

Agenda

- Abstract
- Resumen metadata
- Preguntas de hipótesis a responder
- Respuestas a preguntas de hipótesis
- Insights

Abstract

Este proyecto tiene como objetivo crear un sistema de recomendación de películas mediante técnicas de machine learning.

La principal motivación para el desarrollo del proyecto es explorar sistemas de recomendación, dado que pueden tener aplicaciones importantes en las industrias más operativas, por ejemplo, un recomendador de qué flujo activar en cierto momento, de qué acción puede realizar un usuario en un software, entre otras.

La audiencia del proyecto es principalmente personas que trabajen en áreas de analytics y que busquen explorar un modelo de recomendaciones.

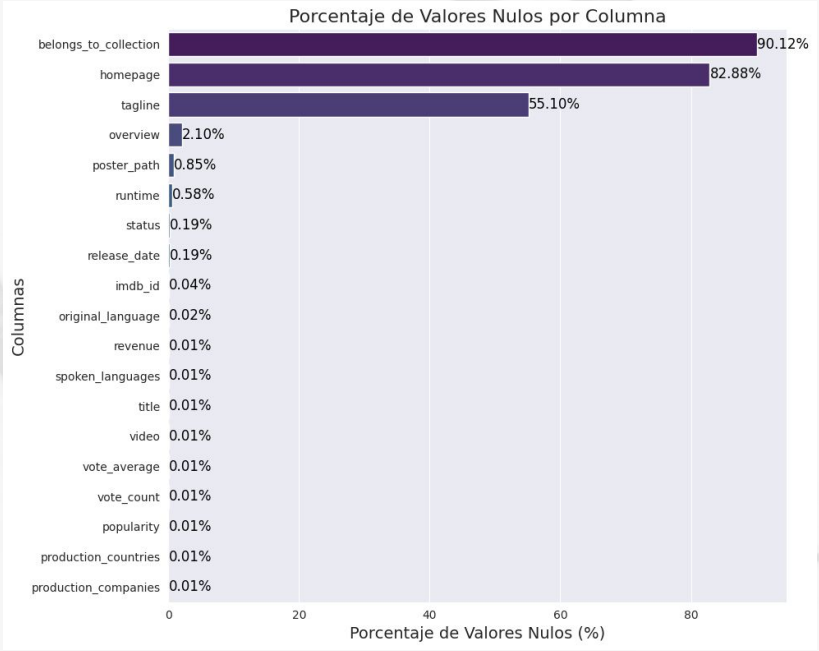
Resumen metadata

A continuación veremos un resumen del estado final de la data procesada

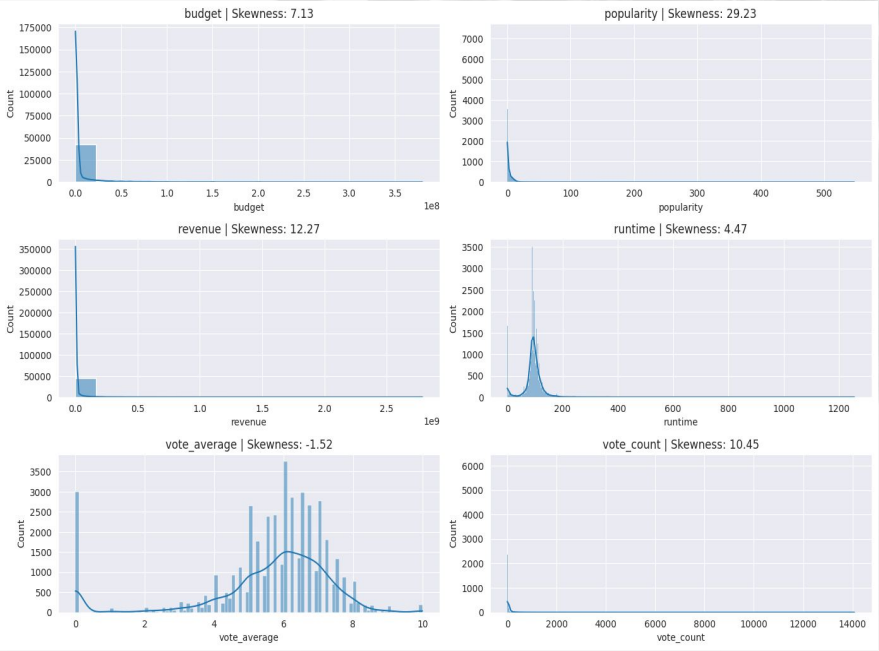
- Cantidad de datos: El dataset cuenta con más de 1 millón de datos (un total de 45.271 registros y 24 columnas)
- Las columnas se dividen en 6 columnas numéricas y 18 columnas categóricas
- De las variables analizadas 19 tenían datos nulos, los cuáles fueron trabajados a la medida de cada una de las variables.
- De las variables numéricas analizadas todas tenían una amplia cantidad de outliers, los cuáles se trabajaron acorde a las necesidades de cada variables

Resumen metadata

Porcentaje de valores nulos por variables



Distribuciones de variables numéricas



Preguntas de hipótesis a responder

Para el proyecto se definieron 3 preguntas de hipótesis a responder

1. Las películas con mejores calificaciones y popularidad tienden a recibir recomendaciones más positivas
2. Los usuarios disfrutan más en películas que comparten género o lenguaje con títulos que ya han disfrutado
3. Las recomendaciones pueden ser más precisas y relevantes si se agregan características como los ingresos, el presupuesto y el idioma [1]

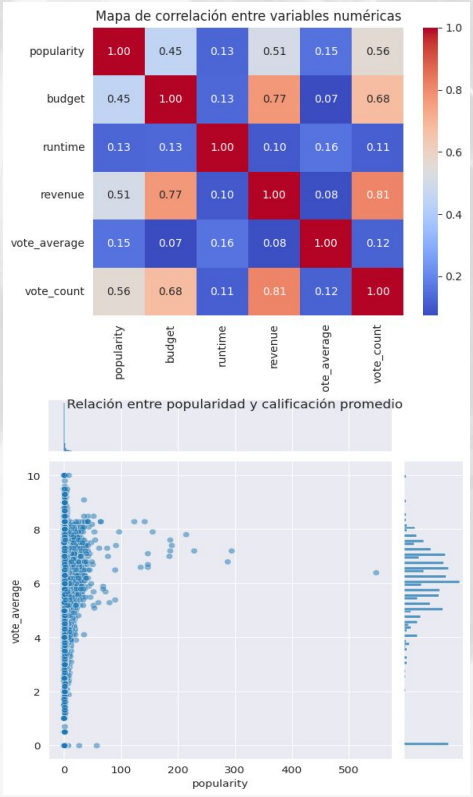
(1) Para poder responder la pregunta 3 se debe generar el modelo recomendador

Respuestas a preguntas de hipótesis

Respecto a la primera pregunta planteada

“Las películas con mejores calificaciones y popularidad tienden a recibir recomendaciones más positivas”

Existe una correlación positiva débil (0.15) entre las películas con mejores calificaciones (vote_average) y su popularidad (popularity). Dicho esto, aunque las películas con mejores calificaciones tienden a ser un poco más populares, no es un factor determinante en si reciben o no recomendaciones más positivas.

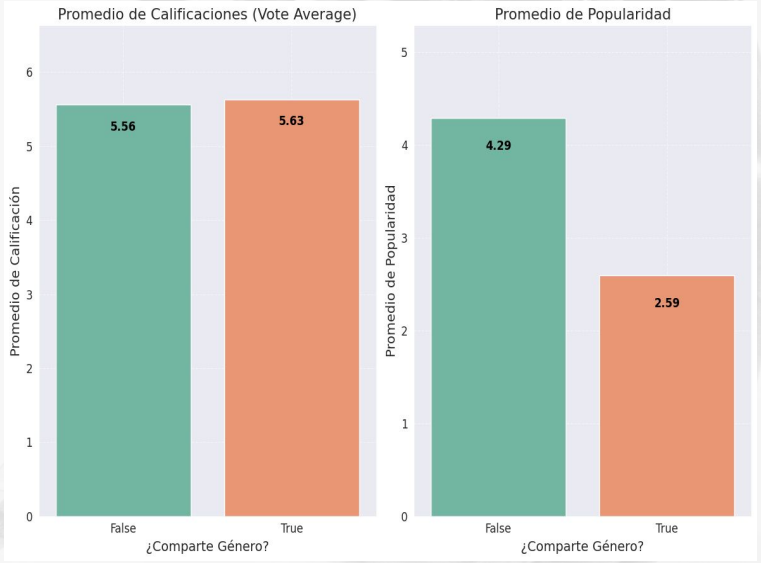


Respuestas a preguntas de hipótesis

Sobre la segunda pregunta planteada

“Los usuarios disfrutan más en películas que comparten género o lenguaje con títulos que ya han disfrutado”

Vemos que las películas que comparten género reciben una clasificación promedio un poco superior a aquellas que no comparten, sin embargo, es llamativo que las películas que no comparten géneros son más populares que las que comparten. Puede ocurrir que películas de géneros diversos atraigan a más población debido a un efecto de explorar nuevos tipos de películas.



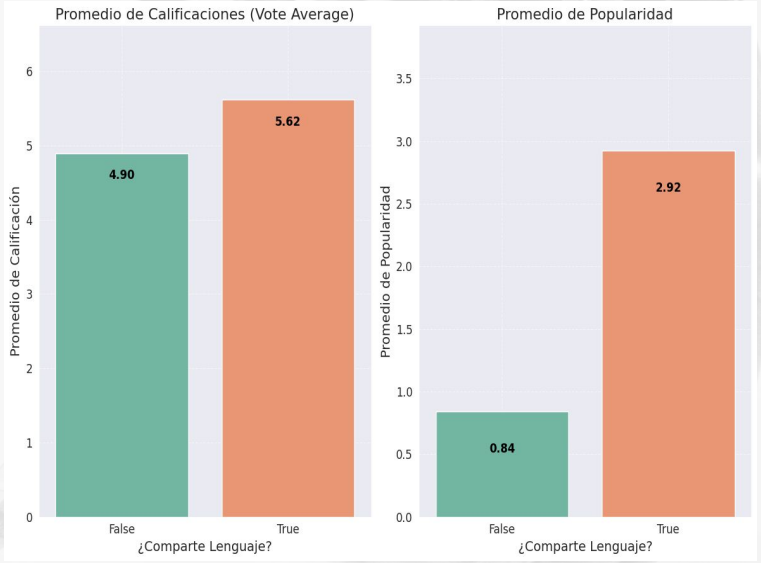
Respuestas a preguntas de hipótesis

Sobre la segunda pregunta planteada

“Los usuarios disfrutan más en películas que comparten género o lenguaje con títulos que ya han disfrutado”

En este caso vemos que en promedio el vote_average si es un tanto mayor que el caso anterior para las películas que comparten un mismo lenguaje original, lo que podría indicar que las personas evalúan mejor películas con un idioma que ya disfrutaron. En el caso de la popularidad si vemos un delta más notable en el promedio en aquellas que comparten un idioma con películas ya disfrutadas.

Esto podría señalar como un factor clave en un sistema de recomendación el lenguaje original



Insights

A continuación presentamos el resumen de los principales hallazgos que surgen del análisis

1. A nivel de correlaciones vemos que las variables de dinero (revenue y budget) se correlacionan fuertemente con variables de popularity (vote_count y popularity), sin embargo, baja correlación para vote_average.
2. La correlación entre las variables vote_average y popularity es baja o débil
3. El género de las películas pareciera no tener incidencia en vote_average de las películas disfrutadas, mientras que sí para la popularity.
4. El lenguaje de las películas pareciera tener una incidencia tanto en vote_average como en popularity para las películas disfrutadas.