Público, crítica y taquilla en IMDb

Análisis exploratorio de datos | Películas 2014 a 2019
Ana Blanco Delgado | Julio de 2021

Objetivo del estudio

Buscar relaciones entre las valoraciones de usuario y de críticos del cine, y las características económicas de las películas como el presupuesto y la recaudación conseguida.

Delimitaciones del estudio

El alcance de este estudio está limitado por la información accesible de manera gratuita en el portal IMDb, y a nivel temporal, se centra en los años comprendidos entre 2.014 y 2.019. El objetivo inicial era estudiar los 10 años anteriores a 2.020 (de 2.010 a 2.019) pero por falta de recursos solo ha sido posible recoger datos para 6 años (2.014 a 2.019).

Fuentes y sus posibles sesgos

IMDb

Desde 1998 pertenece a Amazon, por entonces el objetivo era enriquecer su tienda online de DVDs y cintas de vídeo. Ahora es utilizada por Prime Video para mostrar información sobre películas y series en su plataforma de streaming VOD.
Esto debe tenerse ya que podría dar lugar a un sesgo de información si analizamos películas de producción propia de Amazon.

Metacritic

Ha pertenecido a la CBS Corporation desde 2008 hasta 2020. Así que habría que tener en cuenta que para las producciones de esos años de la CBS podría haber cierto sesgo de información. CBS Films participa en la producción de películas que se estrenan en salas, por tanto, habría que vigilar el valor del metascore para las películas producidas por CBS Films entre los años 2998 y 2020.

OECD

Para las tasas de cambio de los presupuestos de las películas se ha utilizado el portal de la Organización para la cooperación y el desarrollo económicos, que nos proporciona la información de la tasa de cambio estimada para cada país y año.

Valoración de las películas

Disponemos de 3 puntuaciones para cada película:

- IMDb Rating (asociado al número de votos para este rating)
- Metascore
- Popularity

IMDb Rating

Asumimos que el rating es un dato que sigue variando, sigue disponible para que el usuario vote. No disponemos del rating que había cuando la película estaba exhibiéndose en las salas, pero en esa época es cuando más votaciones recibe. Es posible que cada película, tras la emisión por televisión, reciba otra ola de votaciones, pero es cierto que al ser las películas más recientes de 2019 estos impactos ya habrán afectado a los datos recogidos ahora, en julio de 2021.

Metascore

Metacritic es un portal web que recopila críticas de películas, series, programas de televisión, videojuegos y libros. Metacritic convierte cada crítica en un porcentaje y hace una media ponderada para tener en cuenta el caché de la publicación. Esto da como resultado el metascore, una puntuación del 0 al 100 para cada producto, en nuestro caso de estudio, para cada película.
El metascore es habitualmente utilizado por los medios como referencia para medir la recepción de la crítica.
En la página de IMDb aparece este índice para cada película y en este estudio lo tomaremos en cuenta como referencia para evaluar la valoración que hace la crítica de las películas.

Popularity

Dato que vamos a descartar porque es muy volátil, se mantiene en cambio constante y, al ser actual, no guarda relación temporal con la recaudación.

Variables económicas

En IMDb tenemos 4 variables económicas:

- Presupuesto
- Recaudación del primer fin de semana en EEUU y Canadá
- Recaudación en EEUU y Canadá
- Recaudación mundial

De estas variables utilizaremos presupuesto y recaudación. Esta decisión ha propiciado prescindir de muchos registros porque solo 1.553 películas tenían estas 4 variables.

Y además crearemos otras dos nuevas, que serán Beneficio y Retorno de la Inversión, con la siguiente información:

$$Beneficio = Recaudación\ mundial - Presupuesto$$

$$ROI = (Recaudación\ mundial - Presupuesto) / Presupuesto$$

Finalmente, las variables económicas con las que trabajaremos para cada película son:

- Presupuesto
- Recaudación (mundial)
- Beneficio
- ROI

Merge features

Los dataframe finales son el resultado de la unión de varias fuentes enlazadas entre sí con el identificador que IMDb asigna a cada película. Concretamente las features para el merge son las siguientes:

Fuente	**tabla**	**columna**	**descripción**
Base de datos descargada de IMDb	title_basics.tsv 8.084.314 registros	tconst	primary key
		originalTitle	título original
		isAdult	
		startYear	
		runtimeMinutes	
		genres	
	title_ratings.tsv 1.171.920 registros	tconst	primary key
		averageRating	puntuación imdb
		numVotes	nº votos imdb rating
Scrapping películas portal IMDb	movies_df_nnnn.csv	imdb_id	primary key
		title	título para España
		countries	países productores
		metascore	puntuación de metacritic
		popularity	

		budget	presupuesto estimado
		gross_us_canada	recaudación EEUU y Canadá
		opening_us_canada	recaudación fin de semana de estreno en EEUU y Canadá
		gross_world	recaudación total mundial

DataFrame movies

El dataframe utilizado finalmente para el análisis tiene la siguiente información:

variable	descripción	origen	type
imdbId	Identificador alfanumérico de película de IMDb	BD IMDb	`object`
year	Año de estreno (YYYY)	BD IMDb	`float64`
spanishTitle	Título en España	web IMDb	`object`
originalTitle	Título original	BD IMDb	`object`
englishTitle	Título en inglés	BD IMDb	`object`
ratingImdb	Puntuación de usuarios de IMDb. Valores entre 1 y 10.	BD IMDb	`float64`
numVotes	Número de votos que tiene el Rating de IMDb	BD IMDb	`float64`
metascore	Puntuación de críticos de cine, estimada por el portal Metacritic. Valores entre 1 y 100.	web IMDb	`float64`
isAdult	0: non-adult title; 1: adult title	BD IMDb	`float64`
certificate	Calificación de edad extrída de la web	web IMDb	`object`
runtimeMinutes	Duración en minutos	BD IMDb	`float64`
genres	Hasta 3 géneros asociados a la película. Separados por coma.	BD IMDb	`object`
directors	Lista de directores separados por coma	web IMDb	`object`

writers	Lista de guionistas separados por coma	web IMDb	object
stars	Lista de actores separados por coma	web IMDb	object
countries	Hasta 3 países de origen. Separados por coma.	web IMDb	object
companies	Hasta 3 productoras. Separadas por coma.	web IMDb	object
awards	Frase *awards* extraída de la web (string)	web IMDb	object
budget	Presupuesto en dólares ($)	web IMDb	float64
grossUsCanada	Recaudación en EEUU y Canadá en dólares ($)	web IMDb	float64
openingUsCanada	Recaudación en EEUU y Canadá, en primer fin de semana, en dólares ($)	web IMDb	float64
grossWorld	Recaudación mundial en dólares ($)	web IMDb	float64
profit	Beneficio en dólares ($)	generado	float64
roi	Retorno de la Inversión en dólares ($)	generado	float64

Herramientas utilizadas en cada proceso

Web scrapping:

- Visual Studio Code
- Python
- Numpy
- Pandas
- Selenium
- Joblib / Parallel
- Logging
- Pickle

Data mining:

- Jupyter Lab
- Python
- Regex
- Pandas
- Numpy

Visualización:

- Streamlit
- Python
- Pandas
- Numpy
- Matplotlib
- Plotly
- Google Slides

Fuentes

IMDb. Datasets:
https://datasets.imdbws.com/
IMDb. Documentación para los datasets:
https://www.imdb.com/interfaces/

OECD. Tasas de cambio principales monedas por año:
https://data.oecd.org/conversion/exchange-rates.htm

Exchange Rates.Tasas de cambio otras monedas por año:
https://www.exchangerates.org.uk/

Google Developers. Listado de coordenadas de países:
https://developers.google.com/public-data/docs/canonical/countries_csv

Posible map for countries after clean and merge dataframes of years:
https://towardsdatascience.com/using-python-to-create-a-world-map-from-a-list-of-country-names-cd7480d03b10