Movie rating distribution analysis

Zimu Su, Metis Business Project

Abstract:

Movie rating is one of the important metrics to measure audience's feedback and has great impact on current movie industry. Rating modeling analysis leveraged by rating distribution can largely conserve the feedback structure, while arithmetic scoring metrics can generate bias for the controversial movies with polarized ratings. This project presents an analysis of movie rating distribution with respect to film making features, such as budget and genre. It is a preliminary step for the following refined regression analysis to predict potential expectation of movie's outcome (popularity, box office revenue, etc.) based on the movie rating distribution which is served as mediator.

Design:

The rating data of each movie is obtained by aggregating the user rating data by each movie. The rating distribution are categorized into the several patterns based on the cumulative rating number within the intervals between the scores of 0-2, 2.5-3, 3.5-4 and 4.5-5:

Comparison of ratings number among intervals	Rating distribution pattern
0-2<2.5-3<3.5-4<4.5-5	High rating
0-2>2.5-3>3.5-4>4.5-5	Low rating
0-2<2.5-3<3.5-4, 3.5-4>4.5-5	Medium high rating
0-2<2.5-3, 2.5-3>3.5-4>4.5-5	Medium low rating
0-2>2.5-3, 0-2>3.5-4,4.5-5>2.5-3, 4.5-5>3.5-4	Polarized rating
0-2>2.5-3, 2.5-3<3.5-4, 3.5-4>4.5-5	Wave rating

The movies genres and budgets are aggregated based on categories above. The percentage of genre in each distribution pattern to its total number (of movies) is computed and compared to other genres within the distribution categories. A tableau dashboard is created for looking up a genre percentage in each rating distribution. The budgets in each distribution category are analyzed by histogram plots.

Data

The data are collected from https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv, which include movie information (e.g. name, genres, budget), user ratings information. The original source is from GroupLens.

Algorithm

Python pandas is firstly employed to aggregate extremely large amount of user data and count the rating numbers for each score. Google sheets is then utilized to filter the movies based on rating numbers for categorization of rating distribution, as well as genres statistics for each distribution. Tableau dashboard is created for visualization and user interaction with data.

Tools

Google sheet, Tableau, Python Pandas.

Communication:

This project is under supervision of Julia Lintern. The presentation is made on Mar 23rd 2022.