Profitability Analysis and Clustering of Movies

Team Data Science - Economists: Di Kang, Estefania Padilla Gonzalez, Fang Li, Jiuqin Wei, Muhammad Usman Taj
Github: https://github.com/Usman3478/Erdos-Project/tree/main

Overview: In the evolving landscape of the film industry, predicting a movie's profitability is a multifaceted challenge that hinges on a variety of factors, ranging from the storyline and cast to marketing strategies and release timing. Conventional methods often rely on simplistic metrics or historical data, failing to capture the complex interplay of variables that influence a movie's financial success. In this project, we embark on a data-driven exploration to redefine how we can predict a movie's profitability. By leveraging the power of clustering algorithms, we aim to uncover hidden patterns and relationships within a comprehensive dataset of movie attributes. This approach allows us to move beyond traditional binary classifications of success or failure, enabling a more nuanced understanding of the diverse outcomes in the film industry.

Stakeholders: Production companies, Producers, Cinema / Theaters Heads, Streaming Platforms (Netflix, Hulu, HBO)

Data: The TMDb (The Movie Database) is a comprehensive movie database that provides information about movies, including details like titles, ratings, release dates, revenue, genres, and much more. This dataset contains a collection of around 1,000,000 movies from the TMDB database. The dataset can be found here.

Model: We employed K-means clustering for our analysis, which is a widely used method for cluster analysis where the aim is to partition a set of objects into K clusters in such a way that the sum of the squared distances between the objects and their assigned cluster mean is minimized. For our analysis, we divide the movies into three main categories (profitable, break-even, loss-making). Then, the model clusters the whole dataset into these categories on the basis of the given features.

Results: Our findings indicate that the number of votes significantly surpasses other factors in our model, with a notably higher score of 12902. In contrast, all other features have much lower scores. The second-highest feature, "popularity," lags behind with a score of 27. This highlights the vote count as the primary variable and the most dependable predictor of a movie's profitability.

Future Iterations: For possible extensions of this project, it will be crucial to incorporate more up-to-date data and consider additional variables. For example, the state of a country's economy and its citizens' consumption patterns could influence their propensity to spend on movie tickets. During times of economic growth, individuals typically have higher disposable incomes, which can lead to increased participation in leisure activities such as moviegoing or travel. Furthermore, the presence of cinemas can have a substantial impact on a movie's financial success. Greater availability of cinema screens offers audiences increased flexibility and selection, ultimately influencing a movie's sustained profitability.