The Popularity of Movies

Frank Errichiello, Tomasz Hulka, Ankit Kumar Singh

Problem Selection

Every year hundreds of movies are released, with only some making it to mainstream theaters.

The problem we set out to solve is to find the best attributes to predict the popularity of a given movie along with finding the best performing model.

Data Sources

- We used **The Movie Database (TMDB)**
- It contains information on **4803** movies released before October 2017
- The database is hosted on Kaggle

Variables found in the dataset

- budget: The budget used to make the movie.
- genre: The genre of the movie.
- homepage: A link to the website of the movie.
- id: This matches movie_id in the first dataset.
- keywords: Keywords related to the movie.
- original_language: The language in which the movie was made.
- original_title: The title of the movie before translation or adaptation.
- overview: Synopsis of the movie.
- popularity: A numeric quantity specifying the movie popularity.
- production_companies: The production house of the movie.
- production_countries: The countries in which the movie was made.
- release_date: The date on which the movie was released.
- revenue: The revenue generated by the movie.
- runtime: The time of the movie in minutes.
- spoken_languages: a list of the languages spoken in the film.
- status: If the movie has been released or rumored to be released
- tagline: Movie's tagline.
- title: Title of the movie.
- vote_average: average ratings the movie received.
- vote_count: the number of votes received.

Data Science Solution

Using the data modeling techniques of linear regression and clustering we will compare the effectiveness of certain variables to project the popularity of a movie.

Preparation of the Data

- Removed columns we wouldn’t be using.
- Removed any rows that had a status of “Post-Production” or “Rumored”.
- Created new columns with counts of specific details like genres or available languages.
- Split the data into training and test sets using the holdout method with an 80% size training set and 20% size test set.

Results of the linear regression models

- Simple linear regression: The “vote_count” variable best predicts the popularity of a movie
- For variables known before the release of movie, "budget", "num_production_companies", and "runtime" were the best predictor variables
- Multiple linear regression: Best performing model came from the variables "vote_count" and "profit".

Results of the clustering models

- The best performing model was K-means clustering
  - Two clusters, using the variables 'profit', 'vote_count', 'runtime'
    - Supervised metric of .143
    - Unsupervised metric of .688.
- K-means clustering always outperformed the Hierarchical clustering with single linkage
- For Hierarchical clustering
  - Supervised metric was almost always 0
  - Suggested that the clusters were not very similar

Conclusion

Both data modeling techniques of linear regression and clustering support that the best predictors for a movie's popularity are the number of ratings it received (vote_count) and the total profit generated (profit).