Film Industry Trends and Success Using IMDb Data Sets

Part 1: Team & Topic & Database

Group 2: Nathan Harris, Mayumi Shimobe

https://github.com/CSPB4502-Group2/FilmIndustryTrendsAndSuccessUsingIMDbDataSet

CSPB 4502 Data Mining Spring 2025

Description

Our project will examine trends and patterns in the film industry using IMDb’s comprehensive datasets.

We aim to answer questions such as:

- How have entertainment genre popularity trends evolved over time?
- What factors most strongly correlate with higher IMDb ratings?
- What factors influence higher box office earnings?
- Can we build a model to predict a movie's imdb rating and/or box office earnings based on pre-release attributes?

By exploring these questions, we hope to uncover insights into the evolving landscape of movies and TV shows.

Prior Work

- Recommendation engines appear to be the most popular use for Movies and TV data sets. An interesting statistic is that 80% of Netflix streaming time originates from a user clicking on a recommended title. These recommendation engines typically work with content-based recommendation, collaborative recommendation, or a hybrid approach. The content-based approach focuses on similarities between the attributes of a movie or television show whereas collaborative approach focuses on the similarities between the behavior of the user and a class of users. IMDb utilizes its own recommendation engine which can be observed when scrolling to the bottom of a given title under "More like this" which leads us to believe they are focused on a content-based approach. Behavior can be analyzed as explicit or implicit, explicit meaning when a user gives a positive rating to a title while implicit is when they binge watch it in one weekend. As far as predicting the success of film at the box office or with average reviews, it seems that there is no consensus on the indicators for predicting a film's success. Some indicators have been combinations of characters, plot complexity, star power, budget, or "buzz," social media chatter on a particular film. While other indicators are external from a movies data set, such as search engine results, social media content and ratings on expert websites. Evaluating previous movie titles is often used to support these more external data indicators.

Prior Work (cont.)

- In addition, utilizing data sets for movies and television shows also dictates the marketing for a new movie or television show. It doesn't matter how good a movie is unless people know about it (typically through marketing). Meanwhile, casting and talent salaries have been heavily influenced by the success of previous work especially from a data standpoint.

- For our project we will be focused on a content-based approach and utilize ratings as a measurement of success. Utilizing a collaborative approach would involve obtaining demographic user data in addition to implicit/explicit behavior. While we found one data set that includes this, it is far below the threshold for minimum data objects.

Datasets

- List of datasets to use
    - title.basics.tsv.gz: Basic information about movies, TV shows, and video games.
    - title.akas.tsv.gz: Alternate names for titles.
    - title.principals.tsv.gz: Principal cast/crew members for each title.
    - title.crew.tsv.gz: Director and writer information for each title.
    - title.ratings.tsv.gz: IMDb ratings and the number of votes for each title.
    - title.genre.tsv.gz: Information about genres associated with each title.
- Where found:
    - IMDb's dataset page (https://datasets.imdbws.com/)
    - Kaggle.com (https://www.kaggle.com/datasets/harios/box-office-data-1984-to-2024-from-boxofficemojo)
- Whether it you have it downloaded (on who's machine)

    The datasets are downloaded and stored on both Nathan's and Mayumi's computers. It is also downloaded to OneDrive shared folder for the team use.

Proposed Work

- Data cleaning:
    - Handle missing data and incomplete records.
    - Eliminate any duplicate entries
    - Validate and standardize data fields such as titles and genres.
- Data preprocessing:
    - Convert categorical data into numerical representations
    - Extract data features like runtime, genres, directors, and ratings.
- Data integration:
    - Integrate box office earnings data with imdb data based on title and release year
    - Create consolidated tables for analysis
- Data analysis:
    - Identify correlations and trends in features
    - Visualize the correlations and relationships

Proposed Work (conti.)

- Predictive Modeling:
- Target variables: IMDb ratings and box office earnings
- Feature Selection:
  - Pre-release attributes
  - Exclude post-release factors
- Approach:
  - Experiment with different types of models (regression-based, decision trees, or other supervised learning techniques)
  - Use training and testing datasets to evaluate model performance

List of Tools

- Programming Languages: Python (using built-in libraries like Pandas, NumPy, and Matplotlib).
- Data Visualization Tools: Tableau for creating interactive visualizations.
- Development Environment: Jupyter Notebook, VSCode, and/or PyCharm.
- Version Control: GitHub for collaboration, code versioning, and project submission, including Moodle and Piazza.

Evaluation

Trend and Correlation Analysis:

- Compare findings to external sources to validate observations
- Assess how well visualizations answer the main research questions

Predictive Model Performance:

- Evaluate model using appropriate metrics
- Compare different model's accuracy and interpretability
- Where possible, compare predictions to real-world results

Conclusion

We are excited to apply data mining techniques to real-world movie and television datasets, gaining hands-on experience that enhances our understanding of key concepts from this course.

Thank you!