Project Proposal

Project Title: Predicting Movie Revenue

Team Members

Name	Email	JHED
Mao Yang	myang112@jh.edu	myang112
Jia Yu Pan	jpan49@jh.edu	jpan49
Rozanne Lim	rlim8@jh.edu	rlim8
Fan Yu	fyu24@jh.edu	fyu24

Objective

The objective of our final project is to predict box office revenue using attributes of movies such as genre, show length, director, and cast. This information will be housed on an RShiny Dashboard. Our hope is that studios and investors are able to use the dashboard to learn about audiences' movie preferences over time, and what attributes are appealing to the general public to guide future productions.

Existing Work

Frangidis P, Georgiou K, Papadopoulos S. Sentiment Analysis on Movie Scripts and Reviews: Utilizing Sentiment Scores in Rating Prediction. Artificial Intelligence Applications and Innovations. 2020 May 6;583:430–8. doi: 10.1007/978-3-030-49161-1_36. PMCID: PMC7256373.

The paper focuses on recent approaches that use movie reviews and sentiment analysis to predict movie ratings. It introduces a new method that combines the emotional analysis of movie scripts and their corresponding reviews. The main hypothesis is that if the emotional experience expressed by reviewers aligns with or diverges from the emotions conveyed in the movie script, this relationship will be reflected in the movie's rating. To test this, the authors collected a dataset of 747 movie scripts and 78,000 reviews, and applied several machine learning algorithms using vector semantics and sentiment analysis techniques. They compared

their model's performance with conventional approaches. The results show that the proposed feature combination achieves notable performance, comparable to traditional methods.

Telang, R., Sivan, L. Movie sentiment and home entertainment revenue. J Cult Econ 49, 301–326 (2025). https://doi.org/10.1007/s10824-025-09533-5

This study investigates how box office reviews affect home entertainment (HE) sales predictions, which are crucial for forecasting movie revenue. Accurate prediction helps film distributors reduce financial risks when licensing to streaming platforms (like Netflix) or creating promotional strategies. The findings reveal that while the average rating or sentiment score of reviews doesn't directly influence HE sales, extremely negative sentiments significantly impact consumer behavior. In particular, consumers are less likely to purchase electronic sell-through (EST) or video-on-demand (VOD) products when reviews contain strong negative emotions. This suggests that audiences are highly sensitive to negative reviews and adjust their purchase decisions accordingly.

Zhang, Z., Meng, Y. & Xiao, D. Prediction techniques of movie box office using neural networks and emotional mining. Sci Rep 14, 21209 (2024). https://doi.org/10.1038/s41598-024-72340-z

This study focuses on improving movie box office prediction by integrating neural networks and emotional mining (sentiment analysis). The authors note that traditional prediction models have limited accuracy due to incomplete selection of influencing factors. They identify 34 variables across 11 categories, such as movie type, release date, creators, first-day box office, and popularity index. Using Word2Vec, they build a thesaurus for movie-related terms, and create an emotional dictionary based on adjectives and verbs with emotional connotations. The TF-IDF algorithm is then used to quantify the emotional scores of movie reviews. The study compares multiple models — Multivariate Linear Regression (MLR), Back-Propagation Neural Network (BPNN), and Convolutional Neural Network (CNN) — with and without emotional features. Results show that incorporating emotional features (from reviews) significantly improves prediction accuracy: MLR: from 63.4% $\rightarrow$ +16.1% improvement CNN: from 71.9% $\rightarrow$ +11.8% improvement. Overall, the research demonstrates that combining emotional sentiment data with machine learning models notably enhances the precision of box office revenue prediction, providing valuable insights for movie industry decision-making.

Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. Journal of Management Information Systems, 33(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

This study aims to predict movie profitability at early stages of production to support investment decisions. By integrating data from multiple sources and applying social network analysis and text mining, the authors developed a model that analyzes who is involved (cast and crew), what the movie is about (genre and content), and when it will be released (timing). Experimental results show that these combined "who–what–when" features significantly improve prediction accuracy compared to traditional methods. The study highlights how data analytics can not only forecast profitability but also recommend optimal production decisions,

such as profit-maximizing cast combinations, demonstrating the value of predictive analytics in the film industry.

Ahmad IS, Abu Bakar A, Yaakub MR, Darwich M (2020), "Sequel movie revenue prediction model based on sentiment analysis". Data Technologies and Applications, Vol. 54 No. 5 pp. 665–683, doi: https://doi.org/10.1108/DTA-10-2019-0180

This study focuses on predicting the box office revenue of movie sequels, an area that has received limited research attention despite sequels' popularity. The authors propose a sentiment analysis–based supervised learning model combined with a missing value imputation method to improve data quality. Using data from previous sequels within a film franchise, the study applies multiple linear regression (MLR), support vector machines (SVM), and multilayer perceptron neural networks (MLP) to forecast the next sequel's revenue. Results show that using information from four previous sequels yields more accurate predictions than using data from fewer films. The proposed model provides practical value for movie producers and investors, helping them assess the commercial potential of future sequels.

Dashtipour, Kia, Mandar Gogate, Ahsan Adeel, Hadi Larijani, and Amir Hussain. 2021. "Sentiment Analysis of Persian Movie Reviews Using Deep Learning" Entropy 23, no. 5: 596. https://doi.org/10.3390/e23050596

This study introduces a deep learning–based sentiment analysis model for Persian-language movie reviews, addressing the gap in non-English sentiment research. The proposed system automatically classifies reviews as positive or negative, using context-aware deep learning methods. Two architectures — Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) — are implemented and compared with traditional machine learning approaches such as Support Vector Machines (SVM), Logistic Regression, and Multilayer Perceptron (MLP). Results show that the LSTM model achieves the highest accuracy, outperforming other models in detecting emotional tone. This work highlights the potential of deep learning in multilingual sentiment analysis, especially for underrepresented languages like Persian.

Movie Box office Prediction via Joint Actor Representations and Social Media Sentiment https://doi.org/10.48550/arXiv.2006.13417

This study proposes a movie box office prediction model that integrates actor network representations with social media sentiment. It addresses limitations in previous research that relied only on basic film metadata, overlooking social connections and online audience sentiment. The model, named FC-GRU-CNN, combines five key factors — film metadata, Sina Weibo sentiment data, actor social network metrics, shortest path features between actors, and artistic contribution. By leveraging the GRU layer's long-term memory and the CNN's feature-mapping ability, the model captures complex dependencies between actors and audience sentiment. Experimental results show that this method achieves a 14% improvement in accuracy over the best existing C-LSTM model, demonstrating that combining social network information with sentiment data enhances the interpretability and predictive performance of box office forecasts.

Lee S, Choeh JY (2018), "The interactive impact of online word-of-mouth and review helpfulness on box office revenue". Management Decision, Vol. 56 No. 4 pp. 849–866, doi: https://doi.org/10.1108/MD-06-2017-0561

This study explores the interactive effects of online word-of-mouth (eWOM) and review helpfulness on movie box office performance, focusing on the Korean market. While many previous studies have examined eWOM factors such as review count and rating, few have analyzed how the helpfulness of reviews influences box office results. Using data from Naver.com and a sample of 2,090 movies, the authors find that when reviews are perceived as helpful, factors like the number of reviews and review length have a stronger impact on box office revenue. Additionally, review rating, extremity, and helpfulness significantly determine which reviews are deemed useful. The findings emphasize that review helpfulness amplifies the influence of eWOM, offering insights for understanding how audiences' perceptions of review credibility affect movie success.

Lee, S., Choeh, J.Y. The impact of online review helpfulness and word of mouth communication on box office performance predictions. Humanit Soc Sci Commun 7, 84 (2020). https://doi.org/10.1057/s41599-020-00578-9

This study investigates how online review helpfulness and electronic word-of-mouth (eWOM) influence box office performance predictions. While previous research has shown that eWOM factors such as review volume and valence affect product sales, little attention has been paid to the role of reviewer helpfulness. Using data from the Korean movie market (sourced from Naver Movies) and applying various business intelligence (BI) and machine learning models — including random forest, boosted decision trees, k-nearest neighbors, and discriminant analysis — the study finds that movies reviewed by more helpful reviewers or containing more useful reviews achieve higher predictive accuracy for box office outcomes. Moreover, review and reviewer helpfulness strengthen the predictive power of eWOM variables, suggesting that helpfulness moderates the relationship between online word-of-mouth and box office revenue.

Project Outline

The goal of this project is to identify variables that are associated with movie popularity, examine how movie preferences have changed over time, and to use sentiment analysis to identify specific words from audience reviews that can predict box office success. The end product will be a Shiny dashboard with descriptive and interactive components depicting the popularity of movies by variables such as genre, cast, director, and movie length. Additionally, the Shiny dashboard will include a predictive component, by using sentiment analysis of reviews and comments on social media platforms for upcoming movies to predict box office performance.

For this project, several stages are involved:

1. [Research] Identifying variables that could potentially be associated with movie popularity
2. [Data collection] Using publicly available APIs and website scraping from sites such as IMDB, and Box Office Mojo

3. [Data cleaning & wrangling] Using R packages and functional/object-oriented programming paradigms learnt in the course to clean and wrangle data
4. [Data analysis] Conduct appropriate data analysis to answer our questions of interest.
5. [Data visualization] Create interactive data visualizations for Shiny dashboard relevant to our questions of interest
6. [Machine learning] Train model on dataset using supervised learning approach to answer our question of interest: can movies' revenue be predicted by genre
7. [Shiny dashboard] Develop Shiny Dashboard page and include components from data analysis and machine learning

Data sources and API access

- IMDB (https://www.imdb.com/) Basic movie information Cast and crew details Ratings and votes Episode information for TV series Alternative titles
- Box Office Mojo (https://www.boxofficemojo.com) Domestic and international breakdown Theater counts and per-theater averages Opening weekend performance Franchise and sequel comparisons Historical box office rankings
- TMDb (The Movie Database) https://developer.themoviedb.org/docs/getting-started Movie details (title, overview, release date, runtime) Financial data (budget, revenue) Cast and crew information Ratings and popularity scores Posters and images Keywords and genres Production companies

TMDb API Wrapper We have gained access to the TMDb API by registering for an API key. In addition, we are using the TMDb API Wrapper R Package to interact with the API from the R IDE. For box office revenue scores, the `rvest` package was used to scrape data from Box Office Mojo. We are planning to join data for revenues from Box Office Mojo and combine it with data from the TMDb API for other attributes such as genre, cast, and director.

Programming Paradigms

Our goal is to make reproducible and relatively reliable predictions of movie popularity/success (measured by box office revenue) given its attributes (genre, director, cast, movie length, etc.). Functional programming was selected because it encourages a functional style of programming, which is conducive to making our code reproducible. Additionally, we will be using R to conduct most, if not all, of the data cleaning and analyses, and while R is not formally a functional programming language, it contains elements of a functional programming language.

Given that our goal is to make predictions of movie popularity, using a machine learning programming paradigm will address that objective. A supervised learning approach will be used where movie revenues will be paired with labels, which include the movie attributes of interest (genre, director, cast, movie length, etc.).

Packages & Software

Software: R & Positron / VSCode / RStudio Packages:

- Shiny packages for dashboard development
- rvest (HTML) for web scraping
- tidyverse (Meta-package including dplyr, ggplot2, tidyr, etc.)
- lubridate (Date/time manipulation)
- stringr (String manipulation)
- httr (HTTP requests)
- httr2 (Modern HTTP client)
- TMDb (TMDb API wrapper)
- purrr (work with functions, part of functional programming paradigm toolkit)

Data Analytic Product

Our data analytic product will be a RShiny dashboard consisting of at least three separate pages. The home page will include details on the purpose of the dashboard, as well as how to navigate the pages. One page of the dashboard will be dedicated to presenting the interactive data analysis, with graphs displaying various movie attributes. The second page of the dashboard will house details about the machine learning process, and findings.

Timeline

- November 17th - Meeting with Professor Stephanie Hicks
- November 21st - Download data, data wrangling, cleaning and visualization
- November 28th - Machine learning algorithms
- December 5th - Compile elements into the dashboard and build dashboard
- December 10th - Presentation slide submission
- December 11th/16th - Presentation
- December 18th - Final Project Write-Up

Tasks

- Mao Yang - Data cleaning & wrangling
- Rozanne Lim - Data analysis & visualization
- Jia Yu Pan - Machine learning algorithm
- Fan Yu - Setting up the shiny R dashboard