Predicting Movie Box Office Success Based on Reviews

Group 1: Semin Ahn, Catherine Choi, Kathryn Hancock | Group leader: Semin Ahn
DS 4002 | January 30, 2025

Narrative Paragraphs: The film industry is highly unpredictable, with some movies exceeding expectations while others fail despite large budgets. Traditional forecasting relies on factors like production cost, star power, and marketing budgets (Genpact, 2023). However, recent studies suggest that public sentiment in online reviews plays a significant role in predicting box office performance (Mishra & Dey, 2018). Websites like IMDb and Rotten Tomatoes host extensive user reviews and ratings, which can provide valuable insights into audience perception. Analyzing these reviews could help studios and investors make data-driven predictions about a movie's commercial success.

This project will focus on extracting IMDb user reviews, Rotten Tomatoes critic and audience scores, and box office revenue data from Box Office Mojo to build a predictive model. Using Natural Language Processing (NLP) and machine learning, we will assess whether sentiment, review length, and linguistic patterns correlate with a film's financial performance. If successful, this model could be a valuable tool for studios looking to predict box office outcomes before release.

Goal Statement: This project aims to develop a machine learning model that predicts box office revenue based on IMDb user reviews, Rotten Tomatoes scores, and ratings.

Research Question: Can sentiment analysis and textual features of IMDb user reviews and Rotten Tomatoes scores be used to predict a movie's box office revenue?

Modeling Approach: We will develop a predictive model using IMDb, Rotten Tomatoes, and Box Office Mojo data. To process textual data, we will apply Natural Language Processing (NLP) techniques such as tokenization, stopword removal, and sentiment scoring using VADER or BERT. For predictive modeling, we selected Random Forest and XGBoost due to their strong performance in handling both numerical and textual data. Random Forest provides high interpretability and robustness to overfitting, making it suitable for structured datasets. XGBoost, a gradient boosting algorithm, offers superior predictive accuracy and computational efficiency (Chen & Guestrin, 2016). Compared to Linear Regression, which assumes a linear relationship and may not capture complex interactions in review data, and Neural Networks, which require extensive computational resources and large datasets, our approach strikes a balance between accuracy, efficiency, and interpretability. Evaluating model performance using R^2 score and Mean Absolute Error (MAE) will ensure the reliability of predictions.

References:

1. Ravid, S. A. (1999). Information, Blockbusters, and Stars: A Study of the Film Industry. *The Journal of Business, 72*(4), 463-492.

2. Mishra, A., & Dey, L. (2018). Predicting Movie Box Office Success: A Sentiment Analysis Approach. *International Journal of Information Management, 43*, 138-151.
3. Hutto, C. J., & Gilbert, E. (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". Proceedings of the International AAAI Conference on Web and Social Media.
4. IMDb Website. Available: https://www.imdb.com/
5. Rotten Tomatoes. Available: https://www.rottentomatoes.com/
6. Box Office Mojo. Available: https://www.boxofficemojo.com/
7. Genpact. (2023). The Evolution of Forecasting Techniques: Traditional Versus Machine Learning Methods. *Genpact Insights.*