Movie Success Prediction and Sentiment Study

1. Introduction

Movie success is multifaceted, with many determinants such as budget, popularity, and reception by the audience. This project utilizes TMDB movie data and IMDB reviews in creating models that predict revenue and sentiment analysis, and it identifies important drivers of success.

2. Abstract

This project forecasts box office revenues from movies based on machine learning regression models applied on TMDB datasets and examines viewer sentiment of IMDB reviews using NLP. Correlating structured data with text analysis delivers insights into movie success drivers and public opinion.

3. Tools Used

Python (Pandas, Scikit-learn, NLTK - VADER), Matplotlib, Seaborn, Jupyter Notebook.

4. Steps in Project

a) Data Preparation: Cleaned and integrated TMDB datasets; log transform on revenue applied.

b) Modeling: Trained Linear and Random Forest regressors to forecast revenue; Random Forest provided higher accuracy (R^2 = 0.61).

c) Sentiment Analysis: Utilized VADER to label IMDB reviews; attained ~69% accuracy against labeled sentiments.

d) Visualization: Generated plots of prediction outcomes, feature importance, sentiment distribution, and word clouds.

5. Conclusion

The project melds regression and NLP to forecast film success and gauge audience sentiment. Budget was the best revenue predictor. VADER sentiment closely correlated with human labels, revealing the effectiveness of this combined approach to data-driven film analysis.

Predictive Model Summary

Objective:

Train regression models to forecast a film's box office earnings as a function of metadata attributes (budget, popularity, runtime, etc.) from the TMDB dataset.

Models Used:

Linear Regression and Random Forest Regressor.

Data Preprocessing:

Combined TMDB movie and credits datasets. Choosen numeric attributes such as: budget, popularity, runtime, vote_average, vote_count. Performed log transformation on revenue to decrease skewness and facilitate model learning. Split data into training (80%) and test (20%).

Results:

Model	R^2 Score (Log Revenue)	MSE (Log Revenue)
Linear Regression	0.28	47.16
Random Forest	0.61	25.33

Random Forest did a much better job, which is an indication of its capacity to learn nonlinear relationships.

Most Important Feature (according to importance):

Budget — had the largest effect on estimated revenue. Followed by popularity, vote_count, and runtime.

Main Finding:

Higher-budget movies that are more popular tend to make more box office revenue. Random Forest did a better job than linear regression by more effectively dealing with data variability and feature interactions.