Movie Success Prediction and Sentiment Study

This project focuses on predicting the success of movies using data-driven techniques and performing sentiment analysis on user reviews. The project leverages IMDB/Kaggle datasets to understand the factors influencing box office performance and audience perception.

Abstract

The aim of this project is to combine machine learning techniques with sentiment analysis to forecast movie success. The model uses features such as budget, cast, genre, director, and historical ratings, while Natural Language Processing (NLP) tools such as VADER are applied to analyze audience reviews. This dual approach provides both quantitative and qualitative insights into what drives a movie's success.

Tools Used

Tools Used: - Python (Pandas, Numpy, Sklearn) - NLTK & VADER for Sentiment Analysis - Excel for initial data exploration - Matplotlib/Seaborn for visualizations

Steps Involved in Building the Project

Steps Involved in Building the Project: 1. Data Collection: IMDB and Kaggle movie datasets were gathered, including budget, revenue, cast, genre, and ratings. 2. Data Cleaning & Preprocessing: Missing values were handled, and categorical features were encoded. 3. Sentiment Analysis: Audience reviews were processed using NLTK and VADER to assign sentiment polarity scores. 4. Feature Engineering: Derived features like popularity index, sentiment average, and genre encoding were created. 5. Model Building: Regression models (Linear Regression, Random Forest) were used to predict box office success. 6. Evaluation: The models were evaluated using RMSE and R^2 scores, while sentiment distribution was visualized to identify audience preferences.

Conclusion

Conclusion: The project demonstrates that combining structured data (budget, genre, cast) with unstructured data (reviews and sentiments) improves the prediction of movie success. The sentiment analysis provided valuable insights into audience preferences, showing a strong correlation between positive reviews and higher revenues. This approach can help producers, distributors, and marketers make informed decisions for future releases.