Movie Success Prediction and Sentiment Study

Introduction

The film industry heavily relies on public reception to predict a movie’s commercial success. Viewer sentiment and basic movie metadata can reveal strong patterns that influence box office revenue. This project focuses on using such data to build a regression model that can predict box office earnings.

Abstract

The aim of this project was to predict movie success using structured data such as genre, runtime, and number of reviews, along with sentiment scores derived from viewer opinions. The sentiment scores were calculated using VADER, a lexicon-based sentiment analysis tool. A Random Forest Regressor model was trained to estimate box office performance. This model helps in understanding key factors behind successful movies and can support decision-making in production and marketing.

Tools Used

Python Libraries: Pandas, NumPy, Matplotlib, Seaborn
Machine Learning: Scikit-learn (RandomForestRegressor)
Sentiment Analysis: NLTK (VADER)
Jupyter Notebook

Steps Involved in Building the Project

1. Data Preparation
    - A synthetic dataset was created with features like Title, Genre, Runtime, Review Count, and Box Office revenue.
    - Reviews were simulated to calculate sentiment using VADER.
2. Sentiment Analysis
    - SentimentIntensityAnalyzer from NLTK was applied to extract compound sentiment scores from viewer reviews.

3. Feature Engineering
   - Genre was one-hot encoded.
   - Numerical features were kept as-is.
4. Model Building
   - Random Forest Regressor was selected.
   - Dataset was split into training and testing sets (80-20 split).
   - The model was trained and evaluated.
5. Evaluation
   - MAE: 8.56 million USD
   - RMSE: 15.42 million USD
   - Review Count and Sentiment Score were identified as key predictors.
6. Visualization
   - Created bar plots and scatter plots to explore genre-wise sentiment and sentiment vs. revenue.

Conclusion

The model successfully estimates box office revenue using viewer sentiment and basic movie metadata. The combination of sentiment analysis and machine learning shows promising results in predicting movie success. This approach can be extended by incorporating more complex NLP, real review data, and features such as cast or release timing.