Movie Success Prediction and Sentiment Study

Introduction

Predicting movie success has become an increasingly data-driven process. This project aims to estimate a movie's box office revenue based on various features such as rating, vote count, and sentiment analysis derived from viewer descriptions. Using Python-based tools and IMDB/Kaggle datasets, we build a regression model and perform sentiment analysis to identify viewer preferences across genres.

Abstract

The project involves the application of machine learning and natural language processing (NLP) to analyze movie metadata. Using IMDB data, we clean and preprocess the dataset, calculate sentiment scores using the VADER algorithm from NLTK, and apply one-hot encoding on genres. A linear regression model is trained to predict box office revenue, showing an R^2 score of approximately 0.53. We also generate genre-wise sentiment visualizations to interpret viewer reception. This end-to-end workflow demonstrates the potential of combining structured data with text analysis to drive insights.

Tools Used

- Python: Primary programming language used for data analysis and modeling
- Pandas & NumPy: Data handling and numerical computation
- Seaborn & Matplotlib: Visualization of trends and model output
- NLTK (VADER): Lexicon-based sentiment analysis tool
- Scikit-learn: Machine learning library for regression and metrics
- Excel: Output export for data presentation

Steps Involved in Building the Project

1. **Data Collection & Cleaning**: Loaded the IMDB dataset, removed rows missing target variable (revenue), and filled missing values using median or default values.
2. **Sentiment Analysis**: Used VADER sentiment analysis on movie descriptions, treating them as proxy reviews. Scores range from -1 (negative) to +1 (positive).
3. **Feature Engineering**: Converted genres to binary variables via one-hot encoding and selected key

predictors like rating, votes, metascore, and sentiment.

4. **Model Building**: Split the data into training and testing sets (80/20). Used Linear Regression to model the relationship between features and revenue.

5. **Evaluation**: Achieved R^2 score of 0.53 and MSE ~6100. Indicates moderate prediction strength; more features or models could improve accuracy.

6. **Sentiment Visualization**: Analyzed average sentiment per genre and plotted results, helping identify which genres are most positively perceived.

Conclusion

The project effectively demonstrates the use of sentiment analysis and regression modeling to estimate movie success. The insights derived from sentiment scores reveal that genres like Animation and Comedy tend to have higher positivity. While the linear model gives a reasonable prediction, future enhancements could include using deep learning models or integrating user-generated reviews for richer sentiment analysis. This hybrid approach showcases the powerful synergy between NLP and predictive analytics in entertainment and media domains.