MOVIE SUCCESS PREDICTION AND SENTIMENT STUDY

1. Introduction

The success of a movie at the box office is influenced by various factors such as production budget, genre, cast, and especially viewer reception. Understanding this relationship can help producers and marketers plan better. This project explores how sentiment analysis of viewer reviews combined with numerical attributes like budget and genre can predict a movie's commercial performance.

2. Abstract

This project utilizes a dataset from Kaggle containing metadata and reviews for movies. We analyze review sentiments using the VADER sentiment analyzer and build a regression model using LinearRegression to predict a movie's revenue. The sentiment is quantified into positive, neutral, and negative categories, and this is used as a predictor variable alongside budget and genre. Visual analysis is also done to identify sentiment trends across genres.

3. Tools Used

- **Programming Language:** Python
- **Libraries:** Pandas, Seaborn, Matplotlib, Sklearn, NLTK (VADER)
- **Platform:** Jupyter Notebook
- **Dataset Source:** IMDB Movies Dataset , IMDB Movie Reviews

4. Steps Involved in Building the Project

1. Data Preprocessing

- The dataset was loaded using pandas and initial cleaning steps were performed (e.g., dropping nulls, handling formats).
- Only relevant columns like budget, revenue, genre, overview, and crew were selected.

2. Sentiment Analysis

- Used nltk.sentiment.vader.SentimentIntensityAnalyzer to compute sentiment scores from the overview column.

- Generated a compound score and labeled each movie as Positive, Neutral, or Negative.

3. Feature Engineering

- Encoded categorical columns such as genre using one-hot encoding.
- Scaled the budget and used the sentiment scores as input features.

4. Predictive Modeling

- A LinearRegression model was trained on features including budget, genre, and compound sentiment scores.
- Performance metrics such as **R^2 Score** and **Root Mean Squared Error (RMSE)** were used to evaluate the model.

5. Visualization

- Plotted a heatmap and sentiment bar chart using Seaborn to show how different genres align with sentiment.
- Added labels above bars for clarity.

5. Conclusion

The analysis reveals that combining financial indicators like budget with audience sentiments improves prediction of box office performance. The model was able to reasonably estimate revenue, and sentiment trends provided valuable insight into audience preferences. This approach demonstrates the importance of integrating both quantitative data and qualitative feedback in entertainment analytics. With more extensive data, the model's accuracy and utility in real-world applications could be significantly enhanced.