Movie Success Prediction and Sentiment Study

Introduction

This project aims to predict the box office success of movies and analyze viewer sentiment based on genre. Movie success is measured using `gross` revenue data, while sentiment is estimated through a simulated sentiment score in the absence of real review text. The goal is to explore how factors like rating, votes, budget, and sentiment can influence a movie's commercial performance.

Abstract

Using a dataset of 4,000 movies, this study leverages machine learning techniques to model and predict box office revenue. In parallel, sentiment analysis is conducted by assigning synthetic sentiment scores, enabling exploration of genre-based viewer emotions. A linear regression model is used to identify influential features in determining financial success. Visualization of sentiment trends provides insight into audience preferences across genres.

Tools Used

- **Python** – Core programming language
- **Pandas** – Data cleaning and manipulation
- **Scikit-learn** – Machine learning (Linear Regression)
- **Matplotlib & Seaborn** – Data visualization
- **NumPy** – Numerical calculations
- **Google Colab** – Cloud-based development environment

Steps Involved in Building the Project

1. **Data Collection**
   Dataset `movies_updated.csv` with movie title, genre, rating, votes, budget, and gross was used.
2. **Data Cleaning**
   Removed missing values, corrected column names, converted data types, and prepared numerical features.
3. **Sentiment Analysis**
   Since review text was not available, synthetic sentiment scores were generated to simulate user perception.
4. **Feature Engineering**
   Selected features: `budget`, `score`, `votes`, `sentiment_score`.
5. **Model Building**
   A linear regression model was trained to predict movie `gross` using 80% of the data, and tested on the remaining 20%.

6. **Evaluation**
   Model achieved a reasonable R^2 score, showing correlation between movie features and financial success.
7. **Visualization**
   Bar graphs illustrated the average sentiment score by genre, highlighting audience emotion trends across movie types.

Conclusion

The project successfully demonstrates how data-driven techniques can be used to predict movie success and understand sentiment trends. Despite the lack of real user reviews, simulated sentiment analysis provided meaningful insights. The regression model showed that `budget`, `IMDB score`, and `votes` significantly contribute to box office performance. This study provides a foundation for further exploration using real-time sentiment data and advanced models.