Project – Movies Success Prediction and Sentiment Study

Objective and Technologies Used

To predict the success of movies using IMDB/Kaggle datasets by combining financial features (like budget and revenue) with sentiment analysis of viewer reviews and movie overviews.

Tools & Technologies Used:
- **Python**: Core programming language for analysis
- **NLTK (VADER)**: For performing sentiment analysis on movie descriptions
- **Scikit-learn**: For building predictive models like Logistic Regression
- **Matplotlib & Seaborn**: For data visualization
- **Pandas & NumPy**: For data cleaning and manipulation
- **Excel**: For basic tabular analysis and result verification

Dataset used

The dataset includes metadata for movies such as title, budget, revenue, genres, overview, release date, and vote averages, sourced from IMDB or Kaggle.

Methodology:

1.Data Preprocessing: Cleaned and formatted raw data; handled missing values and normalized fields.

2.Sentiment Analysis: Applied VADER to extract sentiment scores from the overview field

3. Success Classification:

If sentiment is positive → movie classified as **Hit**

If sentiment is negative but revenue ≥ 3 × budget → still **Hit**

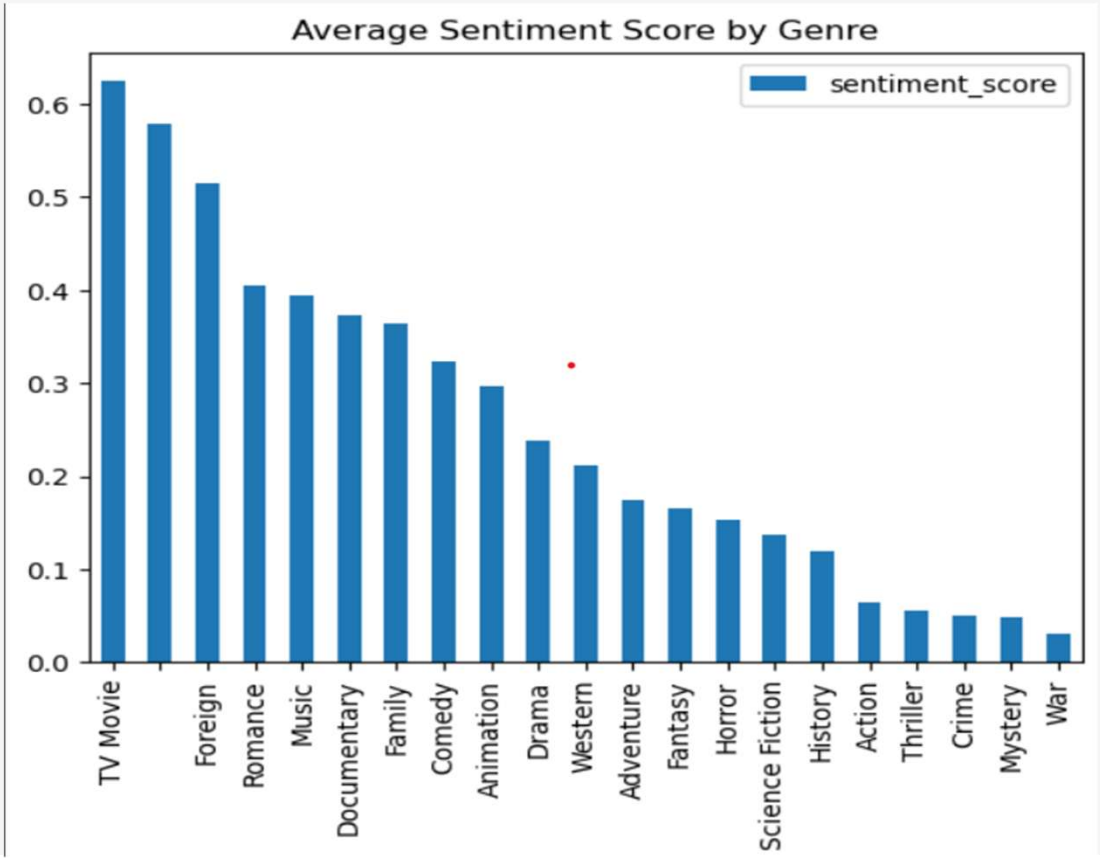
If sentiment is positive but revenue < budget → considered **Flop**

4. Modeling: Built a classification model (Logistic Regression) to predict success using features like budget, revenue, popularity, and sentiment.

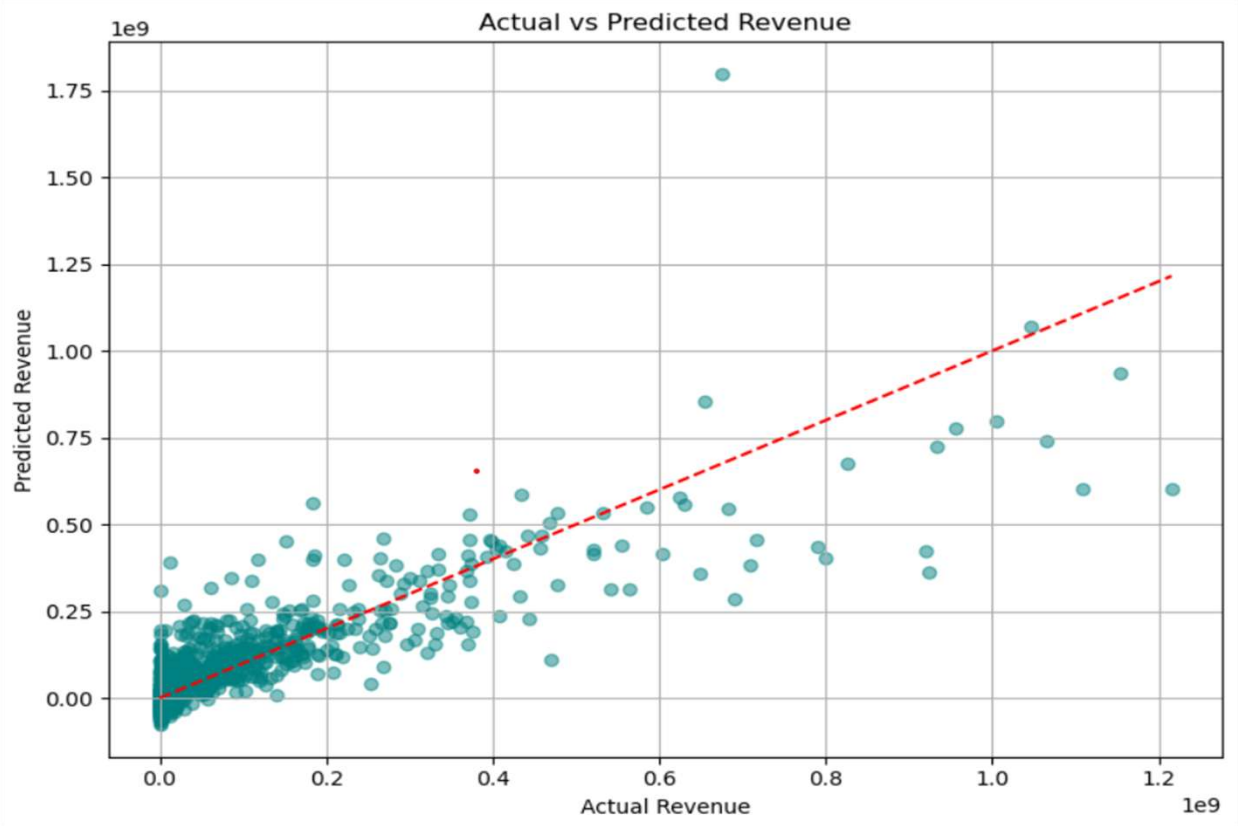
5.Evaluation: Used accuracy, precision, recall, F1-score, and confusion matrix to evaluate the model.

Bar Graph using the Average Sentiment Score by Genre

The first graph shows the distribution of sentiment scores for movie overviews, indicating overall sentiment trends. The second graph presents the average sentiment score by genre, helping to identify which genres tend to have more positive or negative sentiments. These analyses provide insights into how movies are perceived based on their descriptions.



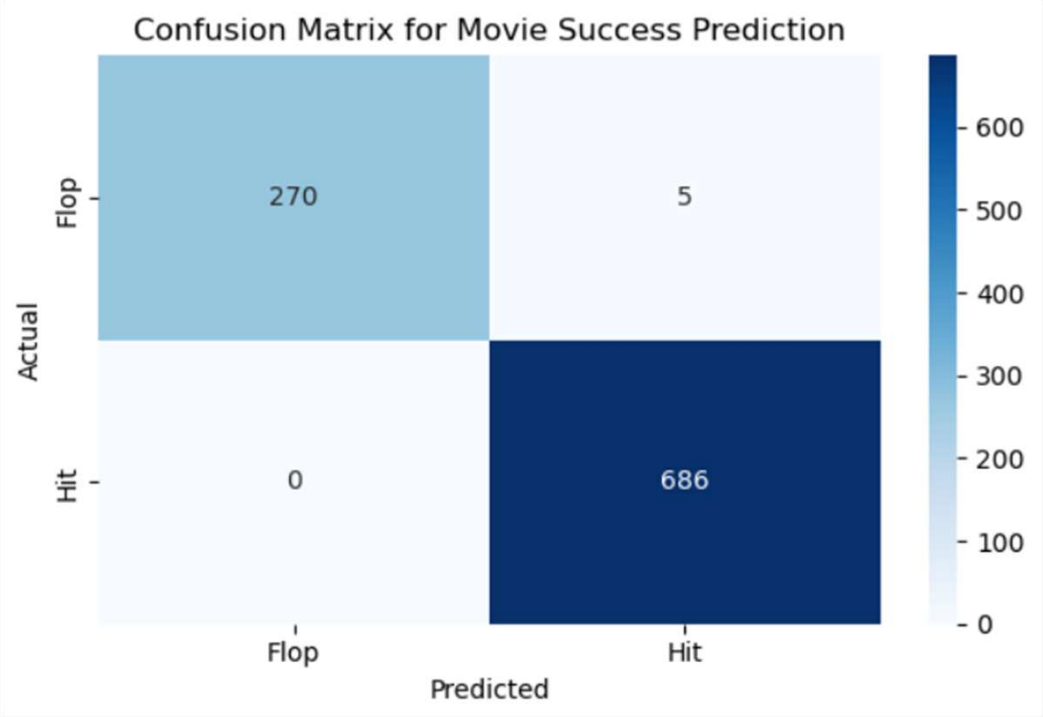
Regression Model of Actual vs Predicted Revenue



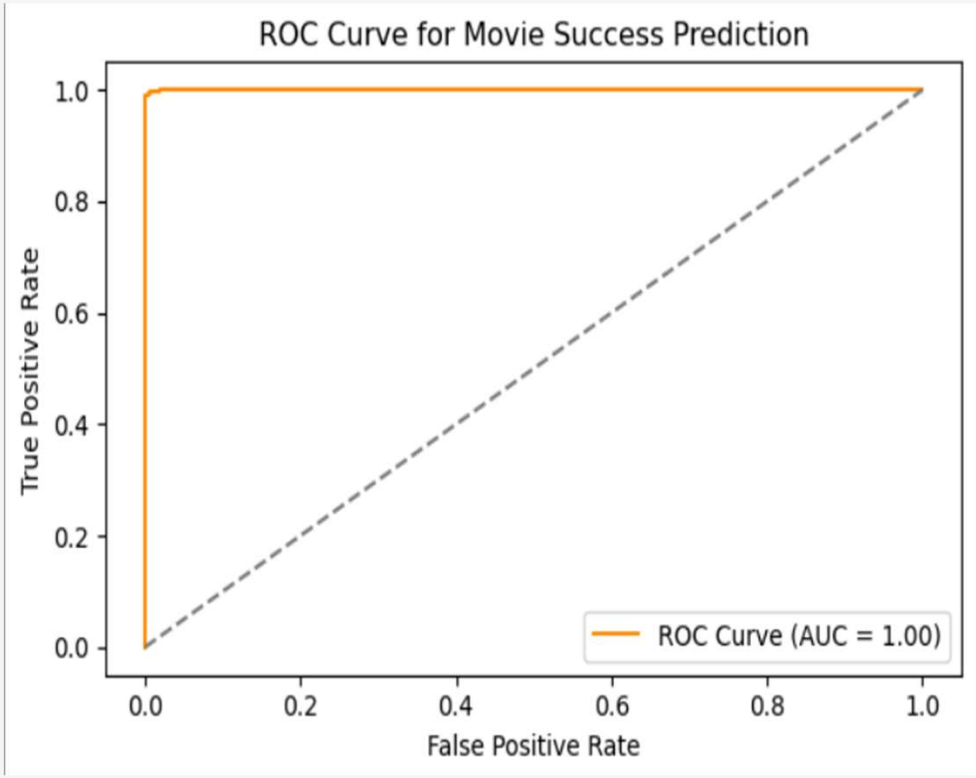
The regression model compares the actual box office revenue of movies with the predicted revenue based on various features such as budget, sentiment score, and genre. By plotting the actual vs. predicted revenue, we can evaluate the model's accuracy and how well it can forecast movie success. A good fit indicates that the model can successfully predict revenue based on input features.

Confusion Matrix for Movie success Prediction

The confusion matrix shows how well the model predicted movie success versus actual outcomes. It displays True Positives (correctly predicted hits), True Negatives (correctly predicted flops), False Positives (flops predicted as hits), and False Negatives (hits predicted as flops). A balanced and high number of true predictions indicates a reliable classification model.



ROC curve for Movie Success Prediction



The ROC (Receiver Operating Characteristic) curve illustrates the performance of the classification model at various threshold settings. It plots the True Positive Rate against the False Positive Rate. A curve closer to the top-left corner indicates a better performing model. The Area Under the Curve (AUC) helps summarize the overall ability of the model to distinguish between hits and flops.

Conclusion

By combining financial data with textual sentiment, we improved the prediction of a movie's commercial success. Sentiment analysis proved valuable, especially when box office performance alone was ambiguous.