Predictive Model Summary

1. Introduction

The predictive modeling section of this project aimed to determine the success of movies based on a combination of metadata features and sentiment scores derived from critic reviews.

2. Models Implemented

- Logistic Regression: Simple baseline classification model, interpretable coefficients.
- Random Forest Classifier: Ensemble method with decision trees, handles complex feature interactions, and provides feature importance ranking.

3. Evaluation Metrics

The following metrics were used for evaluation:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

4. Key Results

- Logistic Regression: Moderate performance, limited in capturing non-linear relationships.
- Random Forest: Achieved higher accuracy and better ROC curve performance. Identified feature importance effectively.

5. Confusion Matrix & Classification Report

Both models were evaluated on confusion matrix heatmaps. Random Forest showed fewer misclassifications compared to Logistic Regression.

6. Feature Importance

Top predictive features identified by Random Forest:

1. Sentiment Score
2. Budget
3. Popularity
4. Runtime
5. Genre Encodings

7. Conclusion

Random Forest was more effective than Logistic Regression in predicting movie success. Sentiment features contributed significantly, proving the importance of review analysis. The model can be further improved with deep learning approaches and inclusion of audience reviews.

Comparative Metrics

Metric	Logistic Regression	Random Forest
Accuracy	Moderate	Higher
Precision	Moderate	Higher
Recall	Moderate	Higher
F1-Score	Moderate	Higher
ROC-AUC	Lower	Better