Comparative Study on ML Algorithms for Predicting Movie Success

1. Introduction

In an era where the film industry is increasingly data-driven, predicting the commercial success of a movie before its release is of immense interest to producers, investors, and marketing teams. This study applies several machine learning algorithms on metadata (cast, crew, budget, genres, etc.) from historical movie data to forecast movie success classes: Flop, Average, Hit, and Blockbuster.

2. Dataset and Features

The dataset contains key movie attributes such as:

- Director and actor names
- Facebook popularity of actors
- Budget, genres, keywords, and duration
- Country and content rating
- Target variable: Success class (0 to 3) based on gross-to-budget ratio

Feature engineering was applied to extract:

- Director and actor success scores
- Genre and keyword features
- Star power (social media influence)
- Release decade and keyword indicators

SMOTE was used to balance the dataset.

3. Algorithms Evaluated

Accuracy Results:

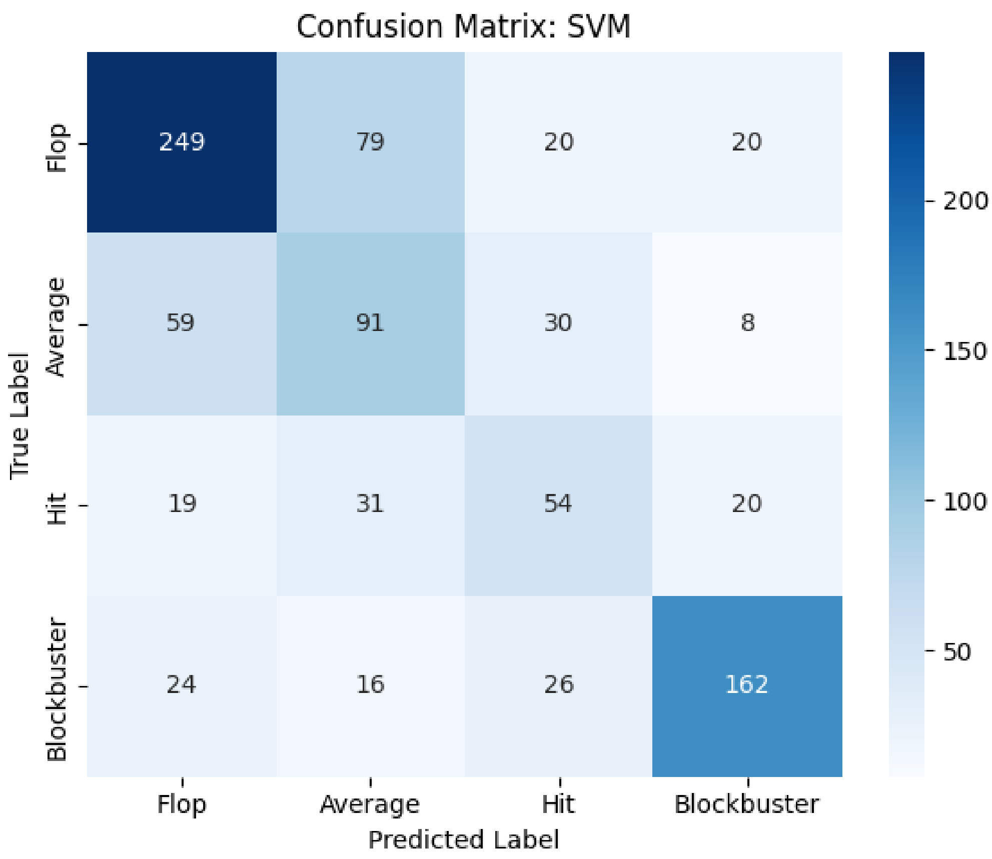
Model	*Accuracy*	***Key points***
SVM	61.2%	Good precision on flops & blockbusters; weaker for middle classes
Random Forest	68.5%	Strong performer overall, especially for class 3
Decision Tree	63.8%	Simple model, prone to overfitting, average across all classes
Logistic Regression	61.6%	Performs surprisingly well despite linear assumptions
Gradient Boosting	**70.8%**	Best overall; excels in capturing non-linear patterns and subtle relationships

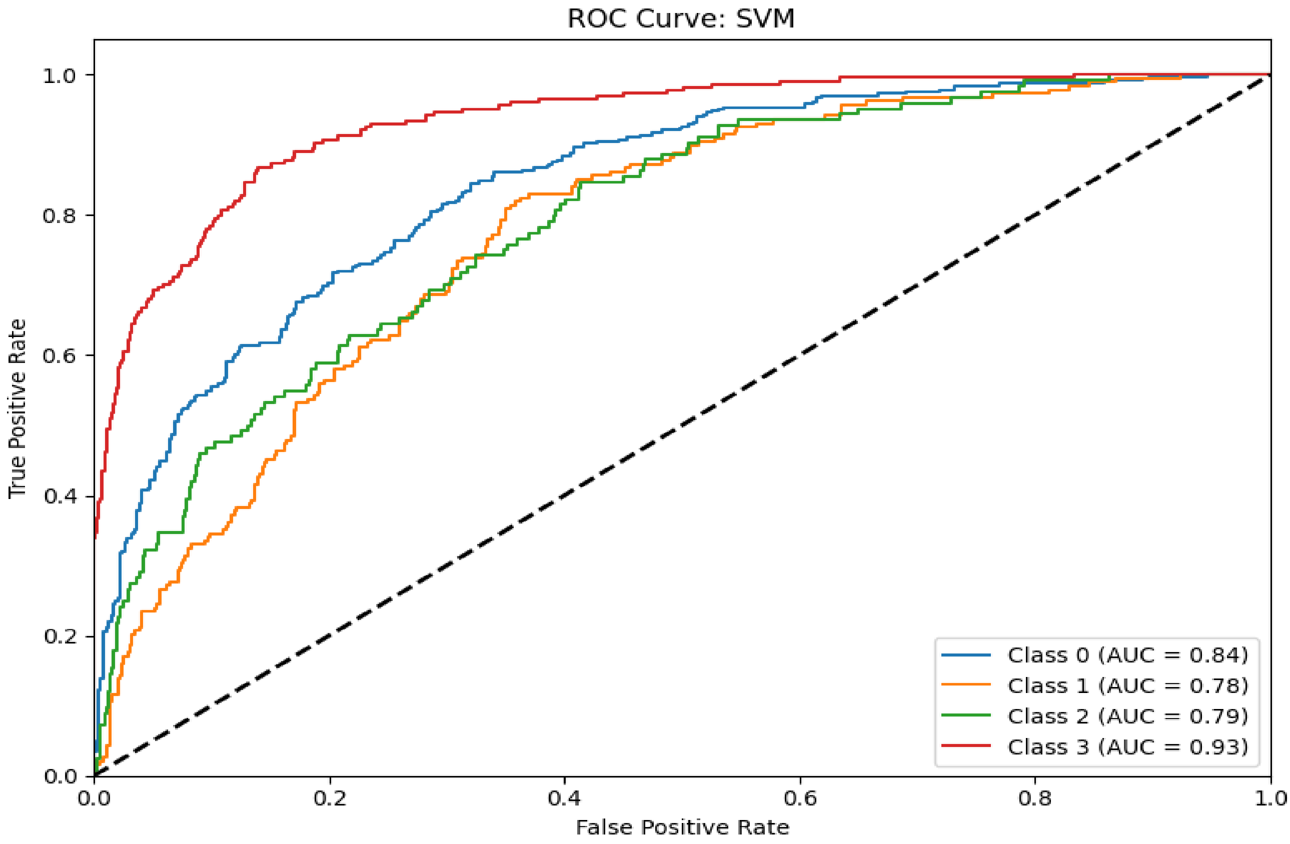
Gradient Boosting outperformed all models, especially in handling class imbalance and capturing non-linear patterns.

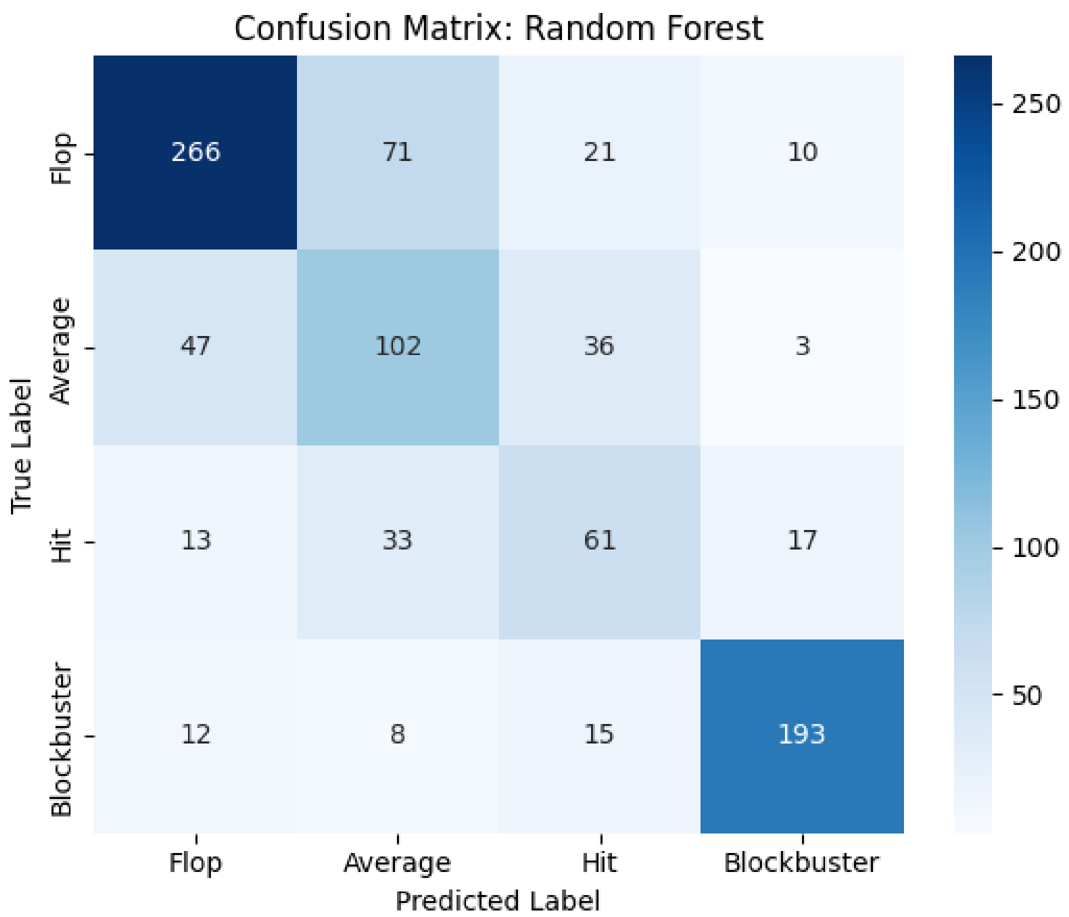
4. Visual Analysis

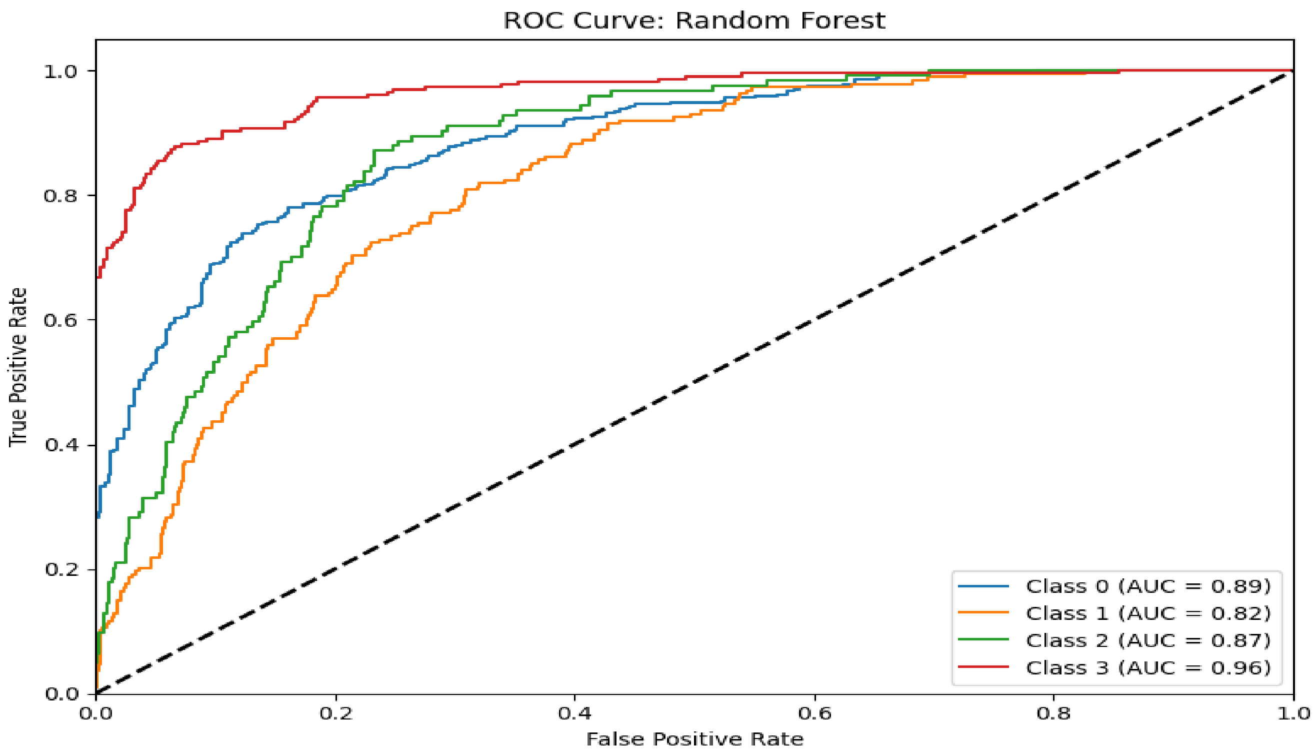
Confusion matrices and ROC curves showed:

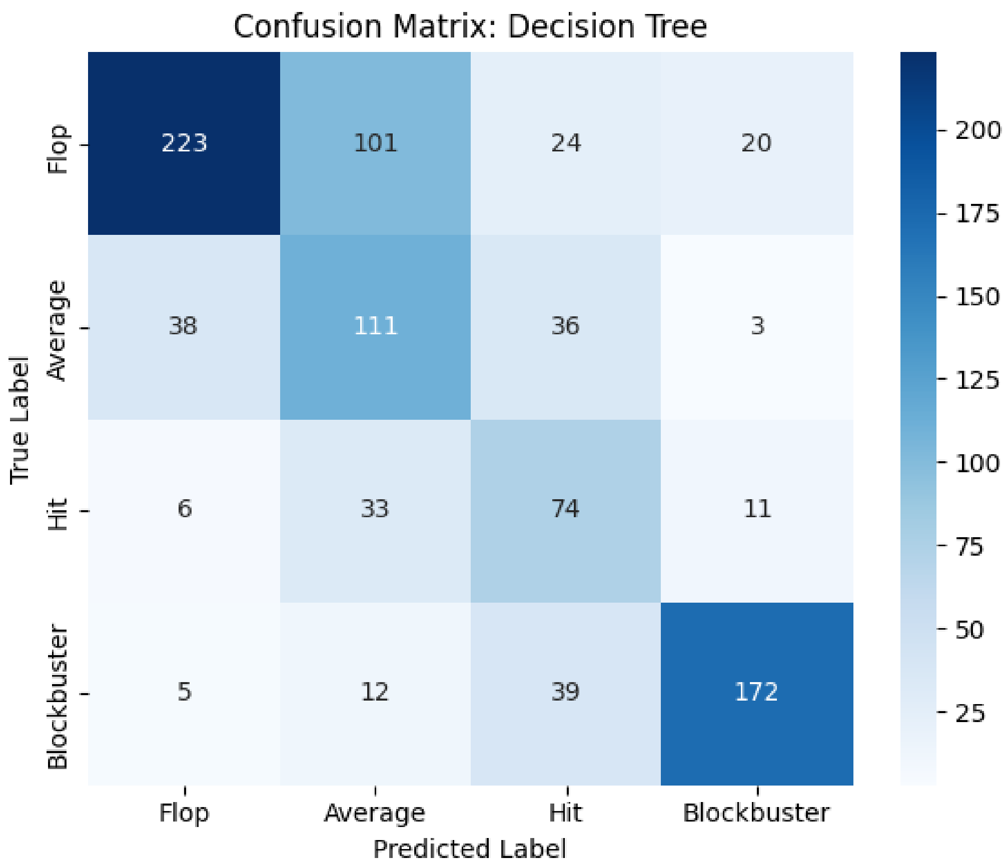
- Strong performance in identifying flops and blockbusters.
- Mixed results in distinguishing average and hit movies.
- Gradient Boosting and Random Forest showed highest ROC AUC scores.

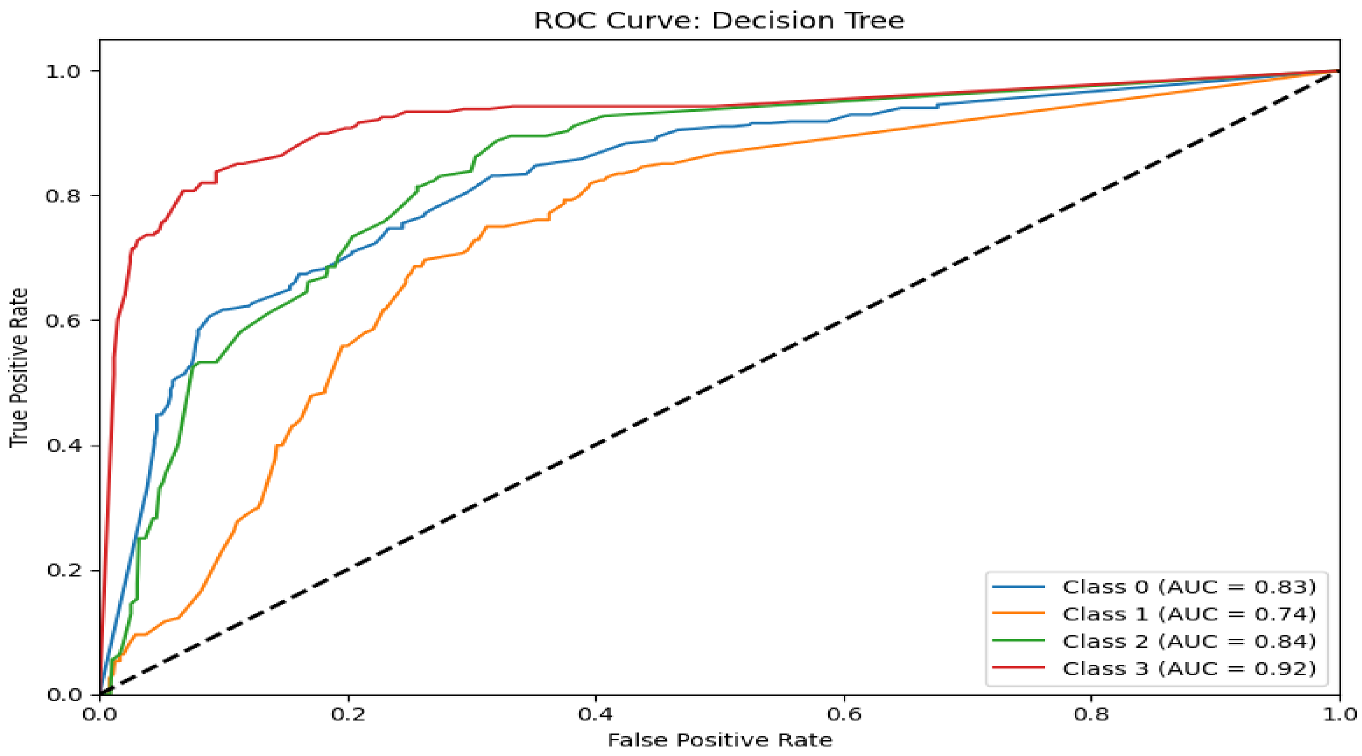


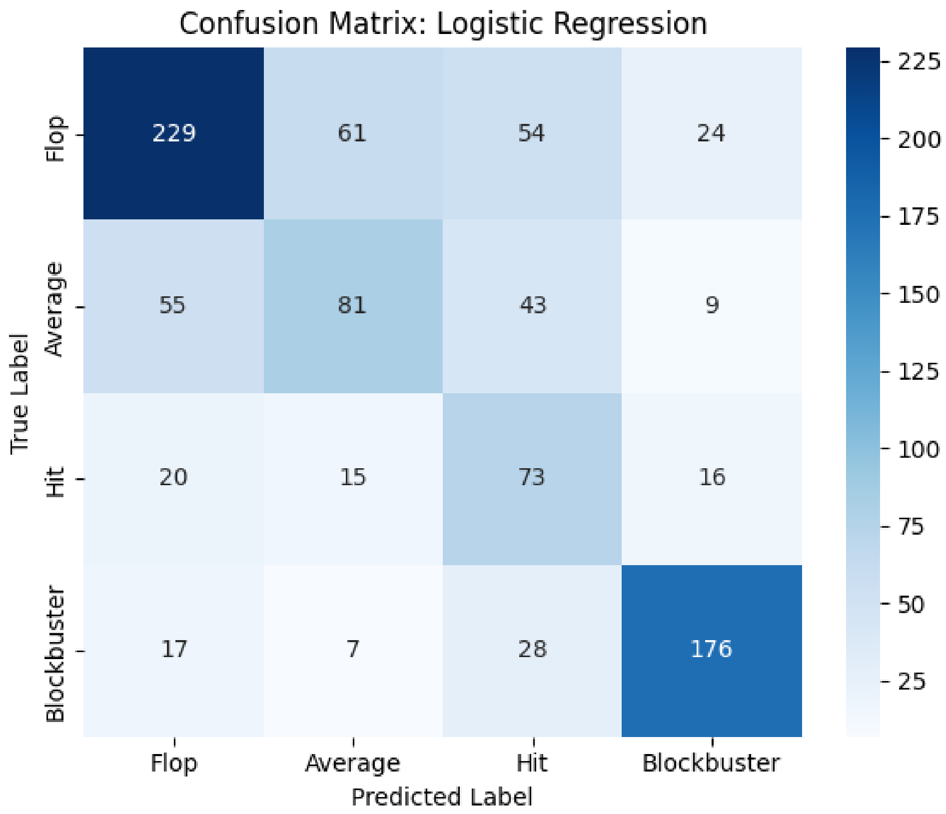

ROC Curve: SVM
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.84)
Class 1 (AUC = 0.78)
Class 2 (AUC = 0.79)
Class 3 (AUC = 0.93)

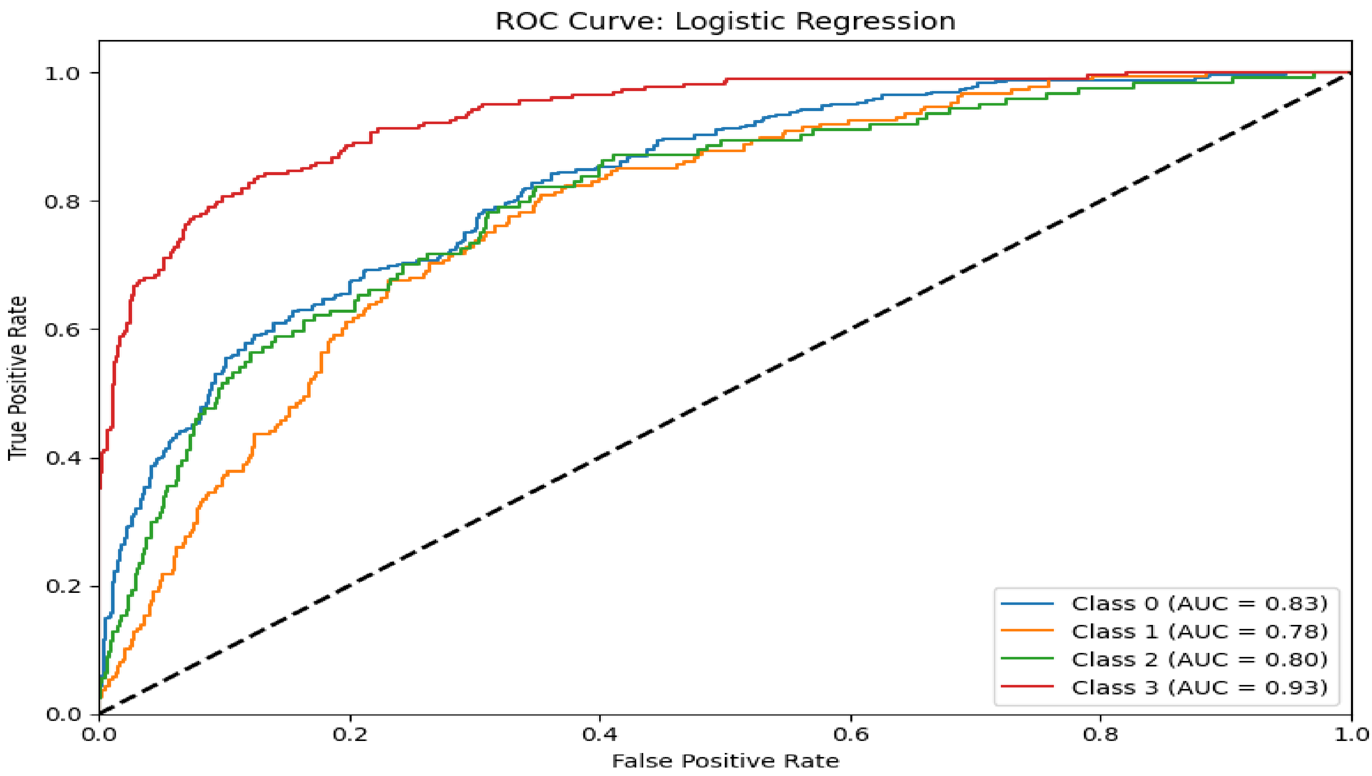

Confusion Matrix: Random Forest
True Label
Predicted Label
Flop
Average
Hit
Blockbuster
266
71
21
10
47
102
36
3
13
33
61
17
12
8
15
193
250
200
150
100
50

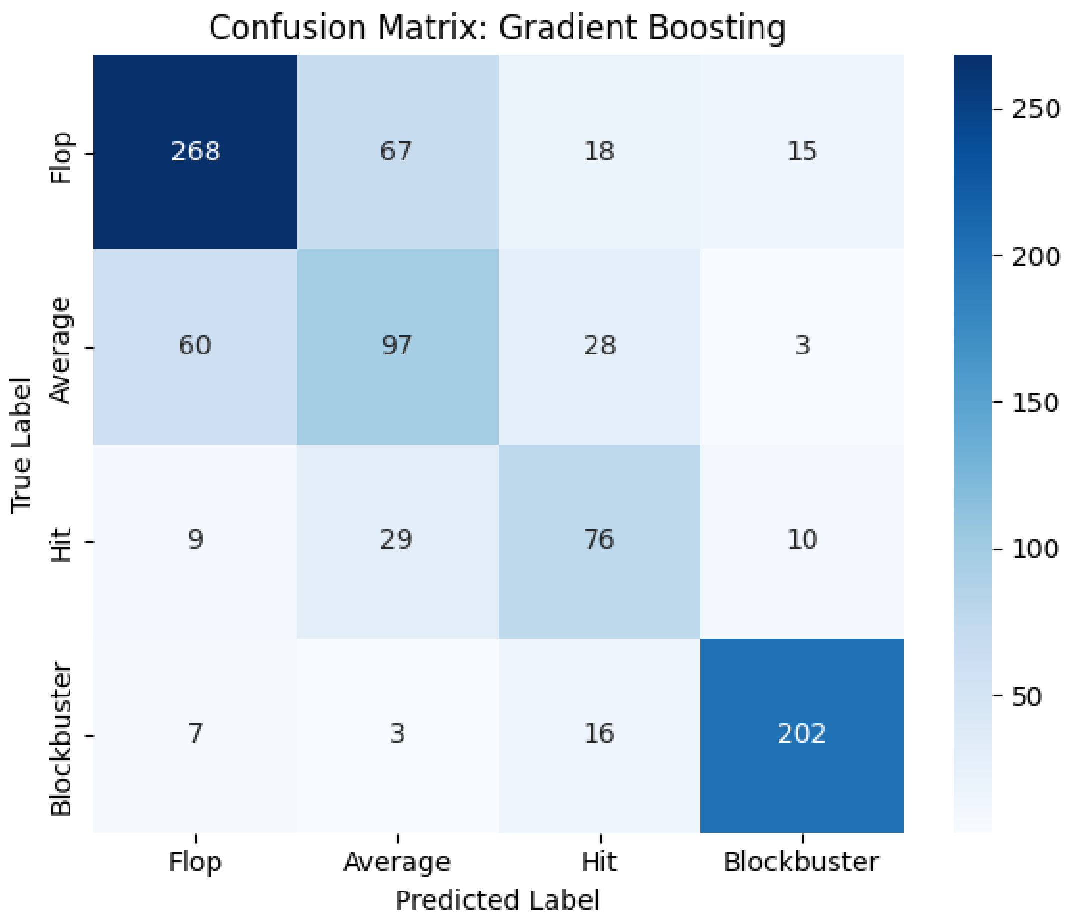

ROC Curve: Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.89)
Class 1 (AUC = 0.82)
Class 2 (AUC = 0.87)
Class 3 (AUC = 0.96)

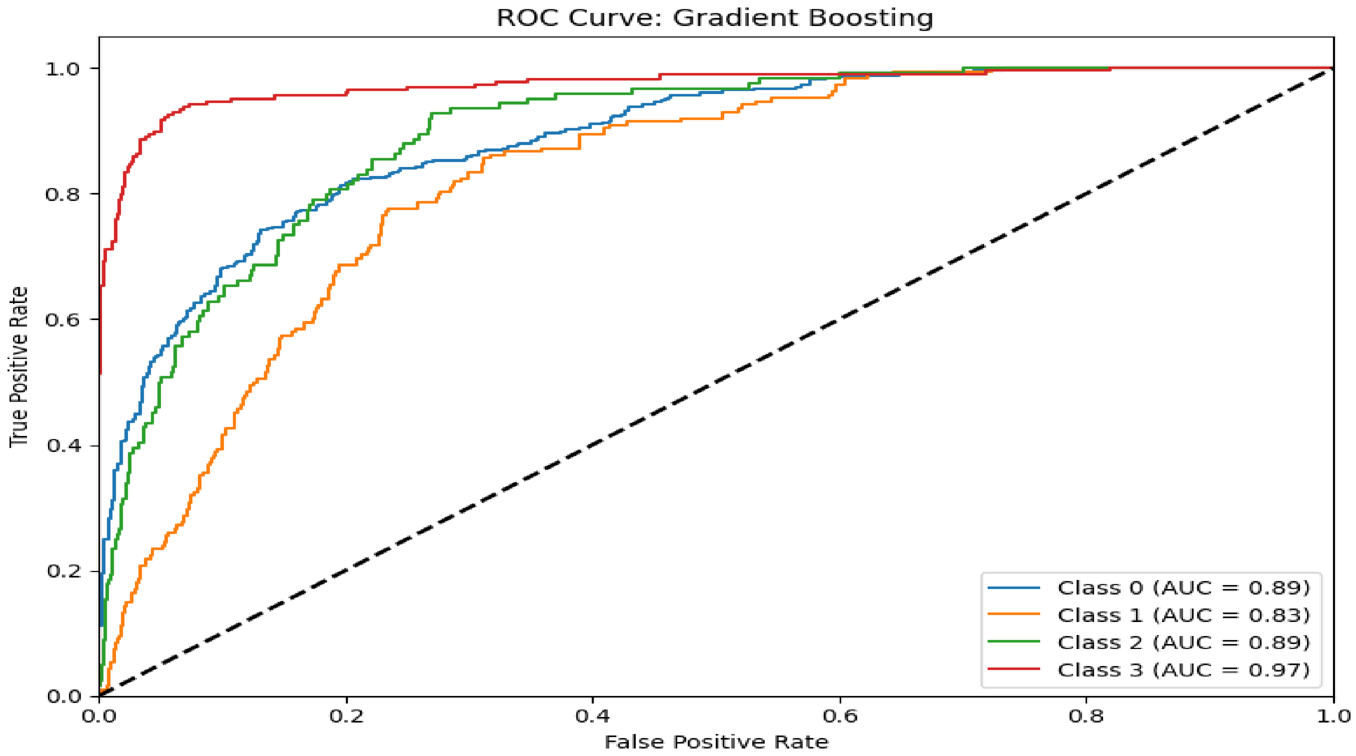

Confusion Matrix: Decision Tree
True Label
Predicted Label
Flop
Average
Hit
Blockbuster
223
101
24
20
38
111
36
3
6
33
74
11
5
12
39
172
200
175
150
125
100
75
50
25


ROC Curve: Decision Tree
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.83)
Class 1 (AUC = 0.74)
Class 2 (AUC = 0.84)
Class 3 (AUC = 0.92)


Confusion Matrix: Logistic Regression
True Label
Predicted Label
Flop
Average
Hit
Blockbuster
229
61
54
24
55
81
43
9
20
15
73
16
17
7
28
176
225
200
175
150
125
100
75
50
25


ROC Curve: Logistic Regression
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.83)
Class 1 (AUC = 0.78)
Class 2 (AUC = 0.80)
Class 3 (AUC = 0.93)


Confusion Matrix: Gradient Boosting
True Label
Predicted Label
Flop
Average
Hit
Blockbuster
268
67
18
15
60
97
28
3
9
29
76
10
7
3
16
202
250
200
150
100
50


ROC Curve: Gradient Boosting
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.89)
Class 1 (AUC = 0.83)
Class 2 (AUC = 0.89)
Class 3 (AUC = 0.97)

5. Real-World Applications

This model can be used in:

- Pre-production risk assessment

- Investor decision-making

- Marketing and promotional budgeting

- Casting and staffing optimization

6. Strengths of the Approach

Strengths include:

- Use of readily available metadata

- Robust to non-linearities

- Effective class balancing with SMOTE

- Scalable to future datasets.

7. Drawbacks and Limitations

Limitations include:

- Subjective success classification

- Missing variables like script quality, market competition

- Overfitting risk in simpler models
- Need for regular retraining to combat future drift

8. Recommendations

To enhance real-world applicability:

- Retrain the model periodically
- Use combined sources like social media data
- Use results as a decision support system and not the accurate outcome.

9. Conclusion

Gradient Boosting was the top performer with ~70.8% accuracy. While perfect prediction isn't possible in the creative industry, these models offer actionable insights for data-informed decision making in film production. If using for financial purposes, additional features are to considered like season, production company, promotion activities and not advised for those purposes yet.