Predicting Movie Success at the Box Office

Ankit Bisht
ankit21014@iiitd.ac.in

Aniket Panchal
aniket21448@iiitd.ac.in

Aryan Sharma
aryan21454@iiitd.ac.in

Syam Sai Santosh Bandi
syam22528@iiitd.ac.in

Aditya Jagadale
aditya22032@iiitd.ac.in

Abstract

The goal of this research is to use multiple machine learning models to predict box office success for films. This originates from a desire to understand and analyse the aspects that influence a movie's box office success, which is classified as "flop," "average," or "hit." Such projections will assist production businesses and investors in making sound judgements based on past film data.

1. Introduction

As we know the box office performance of movies are not influenced by a single variable, It depends on multiple variable ,hence predicting success of a movie is a very challenging task.The above factors involve all these variables, from the budget of the movie's production, star presence, genre, marketing campaigns, release dates, and reviews from the audience. All these can play a major role in whether a film is a blockbuster or a box-office failure or whether it is just average. Predictive, approximate monetary success or failure at the box office is critical for both production companies and distributors, as well as marketing teams, since it may guide important decisions about budgeting, marketing strategies, distribution channels, and even casting.

The purpose of this project is to use machine learning techniques to categorise movies as flops, averages, or hits based on historical data and movie-related criteria. By training machine learning models on these data, the research hopes to create a reliable classification system that can help in predicting the success of new films before they are released. The capacity to classify films as probable flops, averages, or hits can give production firms with significant insights into the aspects that contribute most to a film's success.

2. Literature Survey

Quader, Nahid, Gani, Md, Chaki, Dipankar, and Ali's paper discusses a decision support system that helps investors in the movie industry avoid financial risks by predicting a movie's success based on profitability. Using machine learning techniques like Support Vector Machine (SVM), Neural Networks, and Natural Language Processing, the system analyzes historical data from sources such as IMDb, Rotten Tomatoes, Box Office Mojo, and Metacritic. The authors focus on pre-released and post-released features for the prediction, such as budget, IMDb votes, and the number of screens. They propose a model that classifies movies into five categories, ranging from flop to blockbuster, and give weight to factors like budget and star power are key indicators of success.The study achieves 84.1% accuracy with Neural Networks and 83.44% with SVM for pre-release features, with improved results when considering all features. This research adds to the growing body of work on movie success prediction by integrating both pre- and post-release data, offering investors a more comprehensive and practical tool for decision-making in the high-stakes film industry. [1].

Lee, Park, Kim, and Choi proposed the Cinema Ensemble Model (CEM), a robust machine learning-based approach to predict movie box-office success. This model enhances prediction accuracy by integrating an ensemble of algorithms, including Gradient Tree Boosting, Random Forests, and Logistic Regression. A key innovation in their work is the inclusion of transmedia storytelling—a feature based on leveraging narratives across multiple media platforms, which has been shown to drive greater audience engagement and success. By incorporating pre-release features such as marketing buzz and star power alongside post-release attributes like early box-office performance, the model improves significantly over previous studies, achieving an accuracy of 58.5 %, surpassing earlier models by Sharda and Delen (2006) and Zhang et al. (2009). However, the authors highlight the sensitivity of algorithms like Support Vector Machines to overfitting, particularly in handling pre-release data, suggesting that future research should address these limitations to further enhance prediction accuracy [2].

3. Dataset and Preprocessing

The dataset used for this project was collected from three primary sources: IMDb, TMDb, and Kaggle. Initially, the dataset consisted of 1.1 million entries, with the following attributes:

id, title, vote average, vote count, status, release date, revenue, runtime, adult, budget, imdb id, original language, original title, popularity, genres, production companies, production countries, spoken languages, keywords, directors, writers, primary director

After identifying missing data, we proceeded to clean and preprocess the dataset to ensure its suitability for machine learning models.

3.1. Preprocessing Steps

We cleaned the dataset by:

- Removing null values across key features such as budget and runtime.
- Removing Non-Significant Columns that would not significantly impact the prediction of a movie's success like imdb id, original title, release date, status, id.
- Converting categorical features like genres and production companies into numerical representations using label encoding.
- Creating a target variable by binning the vote average into three categories: Flop, Average, and Hit.

This resulted in a dataset ready for training with machine learning models.

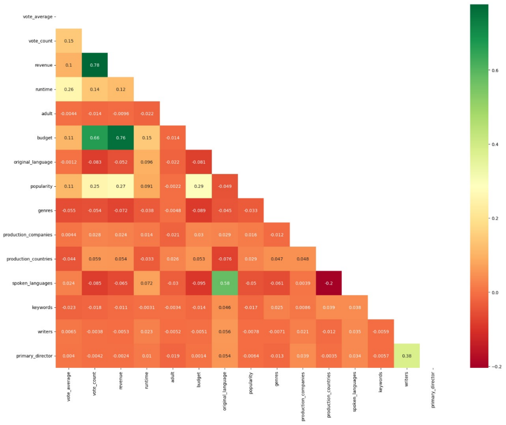

Figure 1. Correlation Heatmap of Movie Features.This heatmap should correlation between different features of our dataset. As we can see there is strong correlation between revenue and vote count, hence to remove multicollinearity we have dropped revenue column.

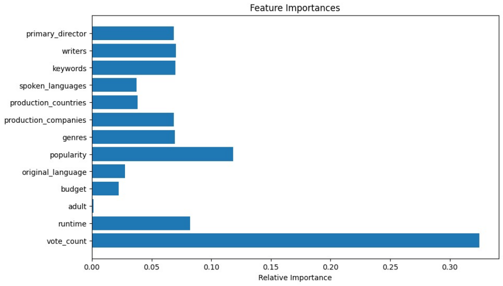

Figure 2. Relative Importance of Movie Features. This bar chart shows the relative importance of movie features in predicting movie success. Vote Count is the most important feature in predicting

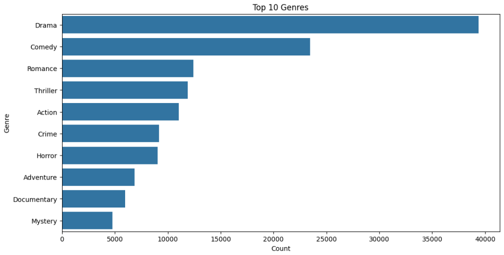

Figure 3. Top 10 movie genres in the data

4. Methodology and Models

In our effort to predict movie success, we experimented with several machine learning models, each model having different strategy to classify the movies into this classes : *Flop*, *Average*, or *Hit*. Below is a detailed explanation of the models tested and their performance:

- **Decision Tree Classifier**: The Decision Tree Classifier is a simple machine learning model that works by splitting the data into subsets based on the most important features. The model creates a tree structure where each internal node represents a feature, each branch represents a decision, and each leaf node represents predicting class. The decision trees are easy to interpret but they leads to overfitting easily especially when the tree becomes overly complex.

  In our case, the Decision Tree has achieved an accuracy of 69%. The accuracy of our decision tree classifier is less than the random forest, decision trees provide useful insights into the structure of the data. The model's F1 score was 0.69, indicating that there is room for improvement in balancing precision and recall. The relatively quick training time of 2.98 seconds makes decision trees a very good option when our priority is computational efficiency.

- **Random Forest Classifier**: Random Forest is an ensemble-based machine learning model that constructs multiple decision trees during training and outputs the class that is the mode of the classifications of the individual trees.It is very effective in handling large number of features as compare decision tree classifier and dealing with non linear relationship between the features. Random Forest reduces the likelihood of overfitting by averaging the predictions of numerous trees, thus creating a more robust and generalized model. In our case, the Random Forest Classifier emerged as the best-performing model, achieving an accuracy of 79% with an F1 score of 0.75. Additionally, the model took 17.12 seconds to train. Random Forest also provides feature importance, which allowed us to identify the most influential factors in predicting movie success (e.g., cast star power, marketing budget, and release timing).

5. Results and Analysis

The Random Forest model achieved the best performance, with an accuracy of 79

- Precision: 77% for Flops, 92% for Average, 68% for Hits.

- Recall: 97% for Flops, 64% for Average, 19% for Hits.

- F1-Score: 0.75 Balancing precision and recall, F1 scores indicate solid performance for Flops and Average, but room for improvement in Hits.

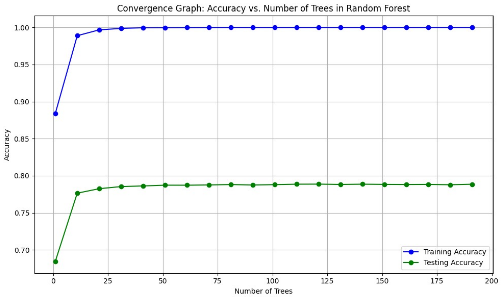

Figure 4. This graph shows the training (blue) and testing (green) accuracy of the Random Forest model as the number of trees increases. The model quickly reaches 100% training accuracy, while testing accuracy plateaus around 79%, suggesting that adding more trees beyond 50 does not significantly improve performance.

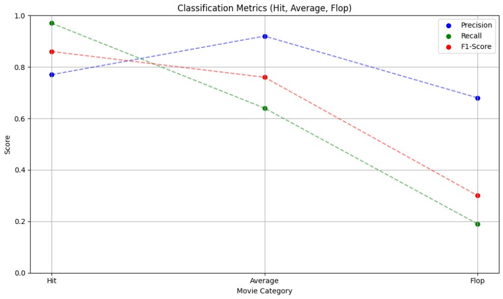

Figure 5. This graph compares precision (blue), recall (green), and F1-score (red) for predicting movie success as Hit, Average, or Flop. The model performs well for "Hits," but recall and F1-score drop significantly for "Flops," indicating difficulty in correctly identifying less successful movies.

6. Conclusion

This project demonstrates that machine learning models, especially ensemble-based methods like Random Forest, can effectively predict box office success. Further improvements could involve training more models on the dataset and tuning them to increase our accuracy.In future we will be exploring boosting based algorithms like XGB classifier and others to increase accuracy of our model.

7. Contribution

Syam and Aditya: Data collection and Exploratory Data Analysis

Aryan and Ankit:Literature survey and model training

Aniket:data transformation and model training

References

[1] T. Quader, S. Nahid, M. Gani, D. Chaki, and M. Ali, "Predicting Box Office Success: An Application of Machine Learn-

ing," *International Journal of Computing and Digital Systems*, vol. 9, no. 2, pp. 139-147, 2020.

[2] K. Lee, S. Park, H. Kim, and J. Choi, ”Cinema Ensemble Model (CEM): Enhancing Box Office Success Prediction Using Transmedia Storytelling," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 86-94, 2020.