e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

MOVIE REVENUE PREDICTION USING MACHINE LEARNING

Tushar Nandkumar Patil*1, Kiran Vijay Desai*2, Prasad T. Goyal*3

*1,2Department Of Computer Science, Shivaji University Kolhapur, India.

*3Assistant Professor, Department Of Computer Science, Shivaji University Kolhapur, India.

ABSTRACT

A lot of variables, including the genre, cast, budget, marketing campaigns, and audience reaction, make it difficult to forecast a film's earnings. This study examines the use of machine learning to forecast a film's revenue by using historical data. Regression models, decision trees, and neural networks are just a few of the machine learning approaches we use to analyse the relationship between a film's numerous components and its box office performance.We use a dataset for this project that contains information about films, such as their budget, cast, genre, director, production firm, release date, and critics' reviews. Important indicators of success including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) are used to train and assess the machine learning models.To improve accuracy, we also use feature selection methods to pinpoint the most important factors that influence a movie's revenue.

Keywords: Box-Office, Machine Learning, Pre-Release, Post-Release Features, Movie Industry, Decision Tree, Random Forest.

I. INTRODUCTION

A key factor in determining profitability in the billions dollar movie business is box office revenue. Movie producers, distributors, investors, and filmmakers have long been interested in accurately predicting a film's box office receipts prior to its debut. In the past, estimates for the box office were made using basic statistical models, market experience, and subjective assessments. But as data from a variety of sources, such as social networking sites, online reviews, and comprehensive production-related indicators, has grown exponentially, machine learning has become a potent tool for producing forecasts that are more precise and grounded in facts.

Because machine learning (ML) approaches provide powerful algorithms that can analyse massive, complicated information and reveal hidden patterns, they have revolutionised many sectors. When it comes to predicting a film's box office performance, machine learning models take into account a number of variables, including the budget, cast popularity, genre, date of release, level of competition, and audience involvement. Machine learning may also measure the audience's excitement before to release by including sentiment analysis from social media and review platforms, which improves the projections' accuracy even more.

The goal of this review paper is to present a thorough analysis of the state of the art in machine learning-based box office prediction research. It draws attention to the benefits and drawbacks of a number of models, such as neural networks, regression analysis, decision trees, support vector machines (SVMs), and deep learning techniques. The study also examines the significance of feature selection, dataset quality, and the influence of outside variables like marketing and public opinion, all of which contribute to the intricacy of predictive modelling.

II. LITERATURE REVIEW

For this project, we created a mathematical model that uses a number of factors to forecast the success or failure of upcoming films. The budget, cast, director, producer, settings, plot writer, day of release, simultaneous release of competing films, release location, music, and target audience were some of the factors used to determine a film's success. Nevertheless, no single trait can be used to predict success. Thus, we developed a model based on intriguing attribute relationships. With the help of this approach, the film business can alter the requirements for movies in order to increase the chance of blockbusters.[1]A tonne of information about the popularity or success of these films may be found online. To generate our data collection, we used the Hollywood movie list from Wikipedia and the ratings from the IMDb movie rating website. From it, machine learning classification techniques were applied to the data set. Ultimately, a useful framework is developed to predict a movie's IMDb rating. When using the data set, the model provides accurate categorization measures.[2]The film business has long been a significant player in the world economy. As a result, it is crucial

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

to forecast the performance of the film before it opens in order to optimize profits.By predicting IMDb ratings for recently released movies using social media data, this study intends to investigate the usage of Factorization Machines approach to movie success and compare it with current studies. Additionally, a framework for compiling movie data from other sources, such as social media, has been established. There are encouraging outcomes when the Factorization Machines are compared to the existing models.[3]We have used different algorithms in the research work for analysis but among all Random forest gave the best prediction accuracy, passing the previous studies. 8 distinct categories are used for evaluation and comparison of the model. A dataset from the real world is used to assess the model. Five distinct assessment measures are used to compare the eight distinct classifiers. The outcomes demonstrate that Random Forest scores better than the remaining categories. Additionally, based on the review criteria outcomes, Ripper Rule Learning did the poorest on the dataset. [4]We are fetching comments and tweets from YouTube and Twitter API respectively after the release of the trailer of the movie and performing Sentiment analysis on them. Therefore, by examining the feelings of film-related remarks on YouTube and Twitter, we present a system that uses machine learning techniques to forecast box office film achievement. [5] A model that combines all of the previous methods is demonstrated. The approach is evaluated and contrasted using 8 distinct classifications. A dataset from the real world is used to assess the model. Five distinct evaluation metrics are used to compare the eight individual classifiers. The results demonstrate that Random Forest performs better than the other classifiers. Furthermore, based on the evaluation metrics results, Ripper Rule Learning did worst on the dataset. [6] The algorithm uses historical data from many sources, notably IMDb, Rotten Tomatoes, Box Office Mojo, and Metacritic, to estimate a film's success rate based on its profitability. Using Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-released features. [7]

III. MACHINE LEARNING MODELS

1. XGBoost

```
1.
XGBoost:

Training Metrics                Test Metrics
R2 score: 0.8633                R2 score: 0.7402
MSLE: 0.0018                    MSLE: 0.0041
```

2. Gradient Boosting

```
2.
Gradient Boosting:

Training Metrics                Test Metrics
R2 score: 0.8750                R2 score: 0.7350
MSLE: 0.0016                    MSLE: 0.0040
```

3. Random Forest

```
3.
Random Forest:

Training Metrics                Test Metrics
R2 score: 0.8475                R2 score: 0.7235
MSLE: 0.0022                    MSLE: 0.0041
```

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

IV. CONCLUSION

The accuracy and efficiency of box office forecasts have significantly increased thanks to machine learning, which uses a range of data sources, including movie information, social media trends, and viewer sentiment. Modern approaches include ensemble methods, transformer-based models, and deep learning outperform outdated statistical techniques in predicting box office revenue in the short and long run. With more and more rich, real-time data being available, future research will focus on enhancing models and incorporating new elements to improve prediction abilities. The film industry will benefit from this smart study.

V. REFERENCES

[1] J Ahmad, P Duraisamy, A Yousef and B Buckles, "Movie success predict ion using data mining", In: 8th International Conference on computing communications and networking technologies ICCCNT 2017, 2017.

[2] W. R. Bristi, Z. Zaman, N. Sultana, Bristi et al., "Predicting IMDb Rating of Movies by Machine Learning Techniques", In: 2019 10th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-5, 2019, 2019.

[3] B. Çizmeci and Ş. G. Ögüdücü, "Predicting imdb ratings of prerelease movies with factorization machines using social media", 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 173-178, 2018.

[4] R. Dhir, A. Raj, Dhir and Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison", In: 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 385-390, 2018, 2018.

[5] T. Sharma, R. Dichwalkar, S. Milkhe and K. Gawande, "Movie Buzz-Movie Success Predict ion System Using Machine Learning Model", In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 111-118, 2020.

[6] Mais Yasen and Sara Tedmori, "Movies Reviews Sentiment Analysis and Classification", IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019.

[7] Quader, Nahid, Md Gani, Chaki Dipankar and Md. Ali, "A Machine Learning Approach to Predict Movie Box-Office Success", 2018.