Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-6 | June - 2025 | Page 72-77

Smart Rating Predictions for Indian Movies Using ML

[1]Nilima Yadav, [2]Vasu Sahu, [3]Mr. Pawan Kumar
[1,2]Students of Amity School of Engineering and Technology, [3]Assistant Professor
[1,2,3]Amity University Chhattisgarh
[1]nilimayadav0726@gmail.com, [2]vasusahupersonal@gmail.com, [3]pkumar@rpr.amity.edu

Abstract

In recent years, the task of predicting movie ratings using machine learning has emerged as a vibrant research domain, drawing attention from both scholars and professionals in the field of data science. This growing interest is largely driven by the surge in publicly available structured datasets, particularly those provided by online platforms like the Internet Movie Database (IMDb). These datasets offer a rich combination of numerical, categorical, and textual attributes related to films, such as cast, crew, duration, genre, and viewer engagement metrics, which collectively enable data-driven insight into audience preferences.

The goal is not only to build predictive systems but also to extract interpretable patterns that contribute to a deeper understanding of what elements drive audience reception. A variety of classification algorithms were employed in this study, all executed within the WEKA machine learning environment. The experimental process included model training, validation, and performance evaluation using standard statistical metrics.

Out of the seven algorithms explored—including decision trees, neural networks, and probabilistic classifiers—the Random Forest classifier exhibited the most reliable results. It achieved an impressive prediction accuracy of 92.7%, outperforming other methods in terms of generalizability and robustness against overfitting. This performance demonstrates the suitability of ensemble methods for handling complex relationships and interactions within film datasets.Beyond technical performance, the results have practical implications for content producers, distributors, and digital streaming services. By accurately forecasting how a film might be received by viewers, stakeholders can make more informed decisions regarding marketing strategies, budget allocations, and release scheduling. Overall, the study reinforces the role of machine learning as a powerful analytical tool in the entertainment industry and opens new possibilities for enhancing audience engagement through data science.

1. Introduction

The global film industry has experienced tremendous growth over the past few decades, not only in terms of the number of films produced but also in the speed and accessibility with which audiences can view them. Thanks to technological advancements and the rise of digital streaming platforms, films are now available to global audiences almost immediately after their release, if not simultaneously. This rapid dissemination of content has led to an overwhelming influx of movie data, highlighting the critical need for effective tools to catalog, manage, and analyze this information.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-6 | June - 2025 | Page 72-77

One of the most valuable repositories for such data is the Internet Movie Database (IMDb), a widely recognized platform that provides structured and unstructured information on movies from around the world. This vast collection of publicly available information has become an essential resource for both the entertainment industry and the data science community.

Leveraging the richness of IMDb's datasets, researchers have conducted numerous studies aiming to forecast movie performance indicators such as popularity, critical reception, and viewer ratings. Different machine learning techniques have been applied to this task with varying levels of success. Algorithms ranging from traditional decision trees to sophisticated neural networks and ensemble models have been employed to explore the relationship between film features and audience response.

In this study, we build upon and enhance previous research by employing a carefully curated and preprocessed IMDb dataset focused specifically on Indian cinema. By cleaning the data, addressing inconsistencies, and normalizing numerical attributes, we created a robust foundation for predictive modeling. This comparative approach not only identifies the most accurate models but also provides insights into which movie attributes most significantly influence viewer ratings, thereby contributing to the broader field of entertainment analytics.

2. Methodology

This study adopted a classification-based machine learning strategy, implemented using the WEKA (Waikato Environment for Knowledge Analysis) platform—a widely used suite for data mining and predictive modeling. Classification, a supervised learning paradigm, involves training algorithms on labeled datasets where each instance is tagged with a known outcome or class. For rigorous evaluation, the dataset was systematically partitioned into training and testing subsets, ensuring that the models were assessed on previously unobserved samples.

A diverse set of seven machine learning algorithms was selected to perform the classification task, each representing a distinct modeling philosophy. These included J48, a decision tree algorithm based on the C4.5 algorithm; Multilayer Perceptron (MLP), which leverages artificial neural networks for pattern recognition; Random Forest, an ensemble method that builds multiple decision trees to enhance robustness; Bagging, another ensemble technique that reduces variance by averaging predictions over multiple learners; BayesNet, which constructs probabilistic models using Bayesian networks; Logistic Model Trees (LMT), combining logistic regression with tree structures for hybrid decision making; and PART, a rulebased learner that generates decision lists by iteratively selecting partial decision trees.

Prior to applying these algorithms, the raw dataset underwent an extensive preprocessing phase to improve data quality and ensure consistency. This involved identifying and handling missing or null values, detecting and correcting outliers, and transforming categorical attributes using encoding techniques. Numerical fields were normalized to a standard scale to prevent features with larger values from dominating the learning process.

Model training and validation were conducted under a 10-fold cross-validation scheme—a robust evaluation technique that splits the dataset into ten subsets, using nine for training and one for testing in each iteration. This cycle is repeated ten times, with each subset serving as the test set once. The results are then averaged to provide an unbiased estimate of model performance. All algorithms were executed using their default configurations within WEKA, enabling a fair

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-6 | June - 2025 | Page 72-77

baseline comparison of their predictive capabilities. The objective of this approach was to identify the most reliable and generalizable model for predicting movie ratings, based on structured metadata extracted from the dataset.

3. Dataset and Preparation

The dataset employed in this research was sourced directly from IMDb's comprehensive and publicly accessible movie database. Initially, the raw dataset encompassed metadata for over 270,000 films spanning various genres, languages, and formats. However, to refine the data for predictive modeling and ensure analytical relevance, the dataset was systematically filtered to include only those movies that featured available information on box office gross revenue—a critical feature for assessing commercial success. To acquire this data, a custom Python script was developed and executed to scrape and extract the "gross" revenue figures from individual movie pages, which were subsequently integrated with the primary dataset through a matching process based on IMDb's unique identifiers.

To ensure data integrity and eliminate redundancy, duplicate entries were identified and removed using IMDb's unique movie codes (commonly known as tconst identifiers). This step was essential for maintaining the consistency and accuracy of the dataset. Additionally, animated films were deliberately excluded to create a more homogeneous dataset, as such films often follow different production, distribution, and reception patterns. After cleaning and filtering, the final curated dataset contained 6,548 unique entries, each representing a distinct non-animated movie.

The final dataset focused on six critical features that were selected based on their potential influence on IMDb ratings. These included:

1. **Gross Revenue** – the reported income from theatrical screenings (usually in the United States),
2. **Start Year** – the year of the movie's initial release,
3. **Runtime (in minutes)** – the total length of the film,
4. **Number of User Votes** – the volume of audience engagement through IMDb voting,
5. **Average IMDb Rating** – the actual score given by users, and
6. **Rating Cluster** – a computed categorical label ranging from 0 to 9 used to discretize the continuous rating into manageable classes for classification.

A crucial aspect of preparing the data for machine learning was **normalization**, which involved scaling the numerical attributes to a standardized range. This process mitigates the risk of features with large value ranges—such as gross revenue or vote counts— dominating the training of the model. To further enhance the learning process, **feature weighting** was applied during normalization. Features historically shown to have a stronger correlation with audience ratings, particularly the number of user votes, were assigned greater significance in the transformation logic. This ensured that the models could better capture meaningful patterns and avoid being misled by scale imbalances or noise in the data.

4. Algorithm Overview J48 (Decision Tree Classifier):

One important feature of J48 is its pruning process, which simplifies the tree by removing parts that do not contribute much to accuracy. This helps the model avoid overfitting, making it more reliable when making predictions on new data. Its decision trees are easy to interpret, which is useful when you want to

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-6 | June - 2025 | Page 72-77

understand how the model makes decisions.

MLP (Multi-Layer Perceptron):
The Multi-Layer Perceptron is a neural network model made up of layers of interconnected nodes called neurons. It is designed to capture complex patterns in data, especially when the relationships are nonlinear. Because of its structure, MLPs can learn from complicated datasets but typically require more processing power and time compared to simpler models. These networks are frequently used in applications where accuracy is a priority, such as image recognition or natural language processing.

Random Forest:
Random Forest is an ensemble learning method that builds many decision trees during training and combines their outputs to make a final prediction. By averaging or voting across these trees, the model reduces the chance of overfitting that a single decision tree might have. This approach improves overall accuracy and stability, making it one of the most popular and reliable algorithms for classification tasks. Random Forests handle noisy data and large numbers of features well.

Bagging (Bootstrap Aggregating): Bagging is a technique that improves prediction stability by training multiple models on different random samples of the data, created by sampling with replacement. Each model makes predictions independently, and the final result is decided by combining these predictions, typically by voting. This method helps reduce variance and makes the model less sensitive to fluctuations in the training data, especially useful for algorithms like decision trees.

BayesNet (Bayesian Network): This allows them to model uncertainty and dependencies between variables in a systematic way. BayesNet models perform well in domains where some features influence others, such as in medical or fault diagnosis systems. While the model assumes some conditional independence among features, it captures complex interactions better than simpler probabilistic models.

LMT (Logistic Model Tree):
LMT combines decision trees with logistic regression by fitting logistic regression models at the leaves of a decision tree. This hybrid design allows it to model linear relationships within specific subsets of data identified by the tree structure. LMTs often provide better predictive performance than standard decision trees and remain interpretable, since the overall decision process is structured like a tree, but with more sophisticated modeling at the endpoints.

PART (Partial Decision Tree):
PART is a rule-based classifier that creates a set of classification rules by repeatedly building smaller decision trees and extracting the best rules from them. Instead of growing a full decision tree, PART focuses on generating a simple, effective rule list. This method strikes a balance between accuracy and ease of interpretation, often resulting in a model that is straightforward to understand while performing well in classification tasks.

5. Results and Analysis

Each machine learning algorithm was evaluated primarily on its classification accuracy. Among them, Random Forest emerged as the top performer, achieving an impressive accuracy of 92.7%. Both Bagging and Logistic Model Trees (LMT) also showed strong results, with accuracy rates surpassing 90%, indicating their robustness in handling the dataset. On the other hand, BayesNet recorded the lowest performance. This lower effectiveness is likely due to its underlying assumption that features are conditionally independent—an assumption that often does not hold true in complex datasets like

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-6 | June - 2025 | Page 72-77

movie ratings, where many factors are interrelated.

The Multi-Layer Perceptron (MLP) showed a more nuanced outcome. It excelled in correctly classifying movies with very high ratings but struggled when it came to movies with average or lower ratings. This pattern suggests that although MLPs are powerful in capturing nonlinear relationships, their success heavily depends on proper tuning of hyperparameters and the availability of sufficient relevant features to generalize well across diverse rating categories.

Analysis of the misclassification trends revealed that most errors occurred in neighboring rating clusters rather than widely off-target predictions. In practical terms, this means that even when the exact rating category was missed, the predicted rating was usually very close to the actual one. This behavior is particularly valuable in rating prediction systems, where small deviations—such as being off by one rating level—are often acceptable and still provide useful insights.

6. Discussion and Future Work

This study demonstrates that machine learning techniques are effective tools for predicting movie ratings, particularly when the dataset is carefully prepared and features are selected with a good understanding of the film industry. The strong performance of Random Forest in this research reflects its well-known ability to manage diverse types of features and handle noisy or complex data efficiently.

Looking forward, there are several promising avenues to improve and extend this work. Incorporating additional features such as movie genre, award recognition (for example, Oscars), official content ratings like MPAA classifications, and sentiment analysis derived from audience reviews could provide richer information for the models to learn from. Additionally, capturing temporal dynamics—such as shifts in audience preferences over time or the influence of economic trends—could enable more precise and context-aware predictions.

Moreover, exploring advanced deep learning approaches or creating ensemble models that blend the strengths of different classifiers might lead to further improvements in accuracy and robustness. Finally, analyzing animation films separately could uncover unique patterns and characteristics distinct from liveaction movies, offering valuable insights for specialized prediction models.

7. Conclusion

The investigation into using machine learning for predicting IMDb movie ratings yielded encouraging outcomes. This study highlights that, when the data is properly prepared and the right algorithms are chosen, high classification accuracy can be attained. Among the models tested, Random Forest, Bagging, and Logistic Model Trees (LMT) demonstrated notable effectiveness, while neural networks like the MultiLayer Perceptron (MLP) showed promise that could be unlocked with additional fine-tuning.

Beyond its academic value, these results hold significant practical potential. Improved rating predictions can enhance movie recommendation engines, inform targeted marketing campaigns, and assist in strategic decisions during film production. As the volume and variety of entertainment data continue to expand, leveraging machine learning techniques will become increasingly important for gaining competitive insights within the film industry.

References

1. Internet Movie Database (IMDb). Available at: https://www.imdb.com/
2. M.H. Latif and H. Afzal, “Prediction of Movies

Popularity Using Machine Learning Techniques," *International Journal of Computer Science and Network Security*, vol. 16, pp. 127–131, 2016.

3. C.D. Butler et al., "Predicting Movie Success Using Machine Learning Algorithms," Proceedings of The Fifteenth LACCEI International Multi-Conference for Engineering, Education Technology, Boca Raton, Florida, USA, 2017.
4. K. Lee, J. Park, I. Kim, and Y. Choi, "Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy," *Information Systems Frontiers*, 2016.
5. T. Yu, "On Predicting the Movie Ratings," Carnegie Mellon University, Human-Computer Interaction Institute, 2017.
6. N. V.R. et al., "Predicting Movie Success Based on IMDb Data," *International Journal of Data Mining and Techniques*, vol. 3, pp. 365–368, 2014.
7. M. Tashman, "The Association Between Film Industry Success and Prior Career History: A Machine Learning Approach," Master's Thesis, Harvard Extension School, 2015.
8. K. Persson, "Predicting Movie Ratings: A Comparative Study on Random Forests and Support Vector Machines," Bachelor Degree Project, University of Skövde, 2015.
9. Rotten Tomatoes. Available at: https://www.rottentomatoes.com/ (Accessed March 26, 2018).
10. MovieLens. Available at: https://movielens.org/
11. E. Frank and I.H. Witten, *Data Mining*, Morgan Kaufmann Publishers, 2000.
12. R. Wirth and H. Jochen Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," 2000.