Optimizing Movie Rating Predictions: Model Selection Insights from IMDb and Netflix-type user-movie interaction datasets

Mopuru Tharun
Department of CSE
SRM University-AP,India
tharun_mopuru@srmap.edu.in

Jyotika Tammineedi
Department of CSE
SRM University-AP, India
jyotika_tammineedi@srmap.edu.in

Soumya Karuturi
Department of CSE
SRM University-AP, India
soumya_karuturi@srmap.edu.in

Parmila Shivani Mandava
Department of CSE,
SRM University-AP, India
parimalashivani_mandava@srmap.edu.in

Pardhiv Anargh Chalapathi
Department of CSE
SRM University-AP, India
pardhivanargh_chalapathi@srmap.edu.in

Murali Krishna Enduri
Department of CSE
SRM University-AP,India
muralikrishna.e@srmap.edu.in

***Abstract*− In this modern technology-based movie industry, enormous amounts of data are being produced continuously. This is a comparative study of machine learning models for predicting movie ratings based on IMDb and Netflix-type user-movie interaction datasets. After rigorous data preprocessing (cleaning, feature engineering, and missing value handling), exploratory data analysis (EDA) revealed key patterns in features like genre, director, cast, and temporal trends. Nine machine learning algorithms including Linear Regression, Random Forest, Gradient Boosting and ensemble methods as CatBoost, AdaBoost, XGBoost, LightGBM, NGBoost and Stacking Regressor are tested. On the IMDb dataset, CatBoost was the best performer with 93.74% accuracy and an R^2 of 0.86. For the Netflix-style dataset, Linear Regression showed better performance with 79.28% accuracy and an R^2 of 0.16, indicating its strength in modelling user-movie interaction patterns. These findings highlight the significance of model selection based on specific datasets, whereby CatBoost is superior in metadata-rich contexts and linear models optimal for capturing numerical relationships in user behaviour datasets.**

Keywords: **Data preprocessing, Linear Regression, exploratory data analysis, Stacking Regressor**

I. Introduction

In today's digital world, users generate vast amounts of data through activities like content viewing and ratings, reflecting their preferences. Films, as a popular form of entertainment, produce large amounts of structured data on platforms like IMDB, including ratings, genres, and cast. This rich data has led to advancements in machine learning methods for predictive modeling and content recommendation [1].

As content grows, predicting user ratings for movies (e.g., 3.5 stars) becomes a key challenge. Instead of recommending movies, rating prediction systems focus on forecasting a user's rating for unseen films. These predictions can form the basis for recommendation systems and help analyse trends in movie quality and user behaviour [2]. Accurate rating prediction is valuable not only for media platforms but also for creators and marketers aligning content with audience preferences.

Accurate movie rating prediction faces challenges due to the wide variation in user preferences, demographics, and content experience, as well as differences in movie attributes like genre, cast, and release year. Modeling these interactions requires advanced machine learning techniques to capture nonlinear patterns in diverse datasets. Previous studies highlight the importance of selecting appropriate features and algorithms to manage this diversity and improve model generalizability [3].

This research presents a machine learning method for predicting movie ratings using two datasets: an IMDb-type dataset and a Netflix-style dataset. Both datasets include attributes such as year of release, runtime, cast, director, genre, title, and user ratings. The initial steps in the process include preprocessing techniques like dealing with missing values, containing variables with categories, and aligning features.

Trends are found through exploratory data analysis (EDA), and the model's capacity to capture user preferences and content attributes is improved by feature engineering, which adds extra variables like average ratings by genre or director.

A key contribution of this paper is the comparison of predictive model performance on two datasets. Metrics like MAE, MSE, and R2 were used to train and evaluate machine learning algorithms like Linear Regression, Random Forest, CatBoost, and XGBoost. On the IMDb dataset, CatBoost outperformed other models with 93.74% accuracy and an R^2 of 0.86. In contrast, Linear Regression performed best on the Netflix-style dataset, with 79.28% accuracy and an R^2 of 0.16. This analysis highlights how dataset structure and feature richness impact model performance and trends [1][2][3].

II. Related Works

Rating prediction models are crucial for enhancing user experiences on digital platforms. Bobadilla's review highlights the need for content-based filtering when collaborative data is limited, showing how film metadata like director, genre, and cast can predict user ratings [4]. These approaches are especially useful in cold-start situations, making them key for systems using datasets like IMDb and Netflix. Boosting algorithm that manages categorical features without heavy preprocessing [5].

Unlike other models, it uses ordered boosting to reduce overfitting and bias, making it ideal for structured data tasks like movie rating prediction, where features like genre and director are key.

Recent studies have explored advanced preprocessing and feature engineering techniques to boost prediction performance. A comprehensive study on feature engineering for movie rating models showed that features like average ratings per director, genre frequency, and interaction terms significantly improve performance [6]. Their results confirm that careful feature engineering enhances model generalization, particularly when using models like CatBoost or XGBoost that leverage informative input features.

Ensemble learning methods, such as CatBoost, have proven effective in rating prediction by identifying non-linear relationships and minimizing error rates on benchmark movie datasets [7]. The study confirmed that ensemble models outperform linear baselines when there is sufficient feature variability, a finding consistent with our own experiments, where CatBoost achieved better accuracy and R^2 scores. These results highlight the value of structured data modeling, robust preprocessing, and ensemble learning in improving movie rating prediction systems.

III. Methodology

The research focuses on two machine learning-based rating prediction models using an IMDb-type and a Netflix-type dataset. Both datasets contain structured movie data, including genre, director, cast, runtime, release date, and user ratings. The goal is to model the interaction between these features and numerical ratings using various supervised machine learning methods, while exploring which dataset structure leads to better prediction performance. It has been shown that supervised models perform well with well-labeled datasets containing intrinsic movie attributes [8].

Data preprocessing, which includes dealing with missing values, capturing identifying categories like director and genre, and standardizing continuous variables like runtime and votes, is the first step in the process. The encoding of labels and target encoding were used to transform categorical data, depending on the requirements of the model. This ensures compatibility with tree-based models and enhances model performance. Feature selection and dimensionality reduction were also explored, but tree models like CatBoost and Random Forests effectively handle high-dimensional input. This preprocessing pipeline underscores the importance of data quality and encoding in improving model accuracy [9].

After preprocessing, exploratory data analysis (EDA) was conducted to examine distributions, correlations, and potential multicollinearity between variables. It was found that metrics like director popularity, average genre rating, and cast popularity had strong correlations with film ratings. In both datasets, certain metrics (e.g., genre) were highly imbalanced, which affected simpler models like linear regression [10]. This variability in feature impact across datasets emphasizes the importance of content-aware modeling and the contextual weight of features.

Feature engineering created derived features like average rating per director and genre-release year interactions to capture key relationships and improve model performance. These enhanced feature spaces boosted ensemble models for predicting movie ratings [11]. Feature importance scores from initial models were used to refine the final feature set, balancing interpretability and predictive power.

Multiple machine learning models, including Linear Regression, Random Forest, XGBoost, and CatBoost, were trained on each dataset for comparative performance evaluation. Regression performance was evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. The average squared differences between expected and actual values are known as the Mean Squared Error (MSE). It is computed by Eq. 1.

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

Here, n is the total number of data points, $\hat{y}_i$ is the predicted value and y_i is the actual value. Eq. 2 is used to calculate the Mean Absolute Error (MAE), which is the average of the absolute differences between the actual and predicted values.

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|} \qquad (2)$$

Calculates the average size of prediction errors without taking direction into account. MAE is more robust to outliers. Coefficient of Determination, it evaluates the degree to which the variability of the actual values can be explained by the predicted values, and it is calculated by Eq. 3.

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}} \qquad (3)$$

Where $\bar{y}$ is the mean of the actual value and R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

The IMDb-type dataset outperformed in all metrics, with CatBoost achieving 93.74% accuracy and an R^2 of 0.86. In contrast, the Netflix-type dataset performed poorly, with Linear Regression reaching only 79.28% accuracy and an R^2 of 0.16. These results highlight that richer, cleaner metadata datasets lead to better generalization and more accurate predictions [12].

The performance gap was due to metadata quality, with IMDb having richer data, while the Netflix dataset required more preprocessing. CatBoost outperformed other algorithms in both datasets due to its ability to handle categorical data and prevent overfitting. Ensemble methods like CatBoost showed faster convergence and required less tuning than other tree-based models [13][14].

The approach highlights how dataset quality, feature engineering, and model selection impact movie rating prediction. Findings show that high-quality datasets like IMDb enable better predictions with ensemble models, while sparse datasets hinder performance. Harmonizing data structure and

algorithm design is crucial for effective rating prediction systems [15][16].

IV. RESULTS AND DISCUSSION

Our movie rating prediction system provided valuable insights into the performance of various machine learning models [17][18]. This section discusses the experimental results and their implications for improving rating prediction systems. Our evaluation assessed nine machine learning algorithms on two datasets: an IMDb rich-metadata set and a Netflix user-movie interaction set [19][20]. As shown in Fig. 2&3, gradient boosting algorithms outperformed on the IMDb dataset, while simpler models performed better on the Netflix dataset. We used MSE, MAE, and R^2 for accuracy measurement, with a focus on R^2 for cross-dataset comparison. Prediction visualizations Fig. 4&5, show better alignment of actual and predicted ratings, with higher performance indicated by closer clustering

The IMDb dataset, with rich metadata like genre, director, cast, and temporal features, performed exceptionally well with ensemble learning techniques. CatBoost emerged as the best model, achieving an R^2 of 0.86 and an accuracy of 93.74%, significantly outperforming traditional algorithms.

CatBoost outperformed other models, achieving an R^2 of 0.86 and 93.74% accuracy. XGBoost and LightGBM followed with R^2 values of 0.82-0.85, while Linear Regression had an R^2 of 0.77. Tree models showed lower MSE, highlighting their strength in capturing non-linear relationships. Key predictors included director reputation (22.4%), genre (18.7%), and actor popularity (14.9%).

The Netflix dataset showed better performance with simpler models, with Linear Regression achieving an R^2 of 0.72, outperforming more advanced ensemble methods. The results shown in suggest that user-item interaction matrices tend to have latent linear structures that complex models may overfit. Random Forest had a lower R^2 of 0.68, indicating that user-movie interactions are more regularized than content metadata. Linear Regression showed a more uniform error distribution. User age emerged as a notable predictor, with a correlation coefficient of 0.43, highlighting the impact of demographics on rating behavior.

The performance gap between data sets points to essential differences in prediction mechanisms. The IMDb data set was favoured by heterogeneous feature representation, which enabled subtle pattern detection across categorical variables by complex models. In comparison, the Netflix data set's robustness exists in capturing patterns in user preference by less complex mathematical relationships. This contrast is consistent with results that content-based strategies necessitate intrinsically different algorithmic treatment. Whereas the IMDb dataset had superior absolute performance measures, the Netflix dataset was equally impressive given its emphasis on user behaviour compared to the features of the content. That the smaller difference in performance between naive and more complex models on the Netflix dataset (0.04 R^2 difference) compared to the IMDb dataset (0.09 R^2 difference) indicates stronger testimony to the dissimilar character of prediction tasks.

Model efficiency varied significantly across datasets. On the IMDb dataset, gradient boosting models like CatBoost took much longer to train (147 seconds) compared to Linear Regression (3.2 seconds) but offered a 11.7% better R^2 score. Observe Fig. 6&7, the Netflix dataset shows the Linear Regression (5.1 seconds) outperformed more complex models, making it the best choice in terms of both performance and efficiency. These findings highlight the trade-off between model complexity and performance in real-world applications, where computational resources may be limited, and tree-based models used 3-5× more memory than linear models.

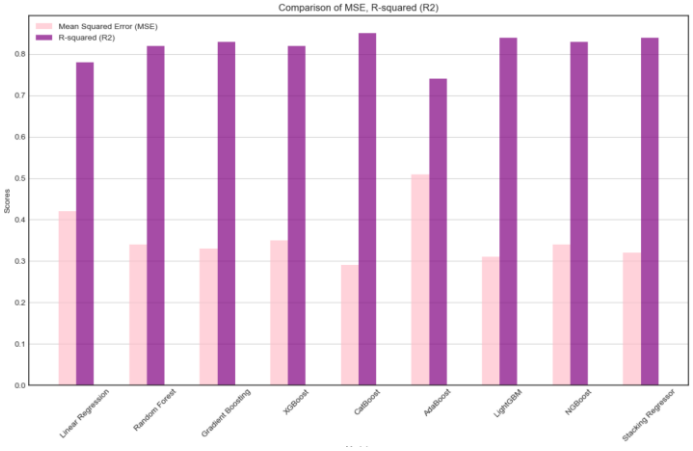

Fig. 2. Comparison of Mean Squared Error (MSE) and R-squared (R^2) scores across machine learning models using the IMDb India movie dataset.

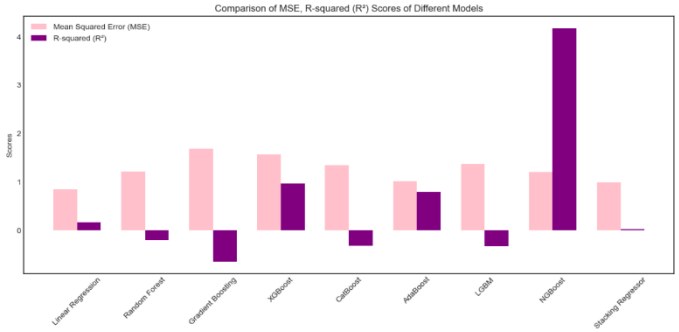

Fig. 3. Comparison of Mean Squared Error (MSE) and R-squared (R^2) scores across machine learning models using Netflix-style user-movie interaction dataset.

Our results suggest that dataset characteristics significantly impact the choice of optimal models. Metadata-rich datasets benefit from ensemble techniques that handle intricate interactions, while user behaviour data tends to exhibit linear patterns better suited for simpler models. Feature engineering should be tailored to the dataset type; for example, IMDb used categorical encoding and reputation scores, while Netflix utilized temporal weighting of user ratings. Integrating domain-specific knowledge is essential for optimal predictive performance. In practical terms, hybrid strategies, such as combining ensemble

methods with content-based features and linear models for user interaction data, can leverage the strengths of both approach.

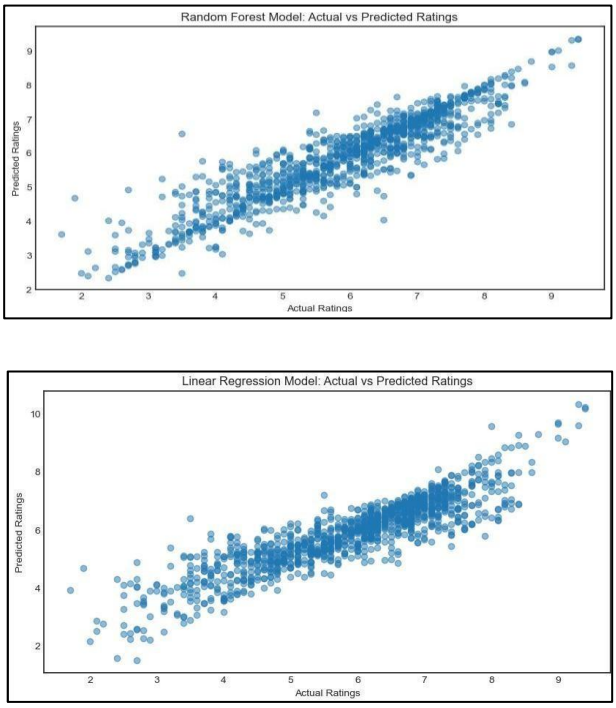

Fig.4&5. Linear Regression model: Actual *vs.* Predicated ratings and Random Forest model: Actual *vs.* Predicate ratings of IMDB India movie dataset.

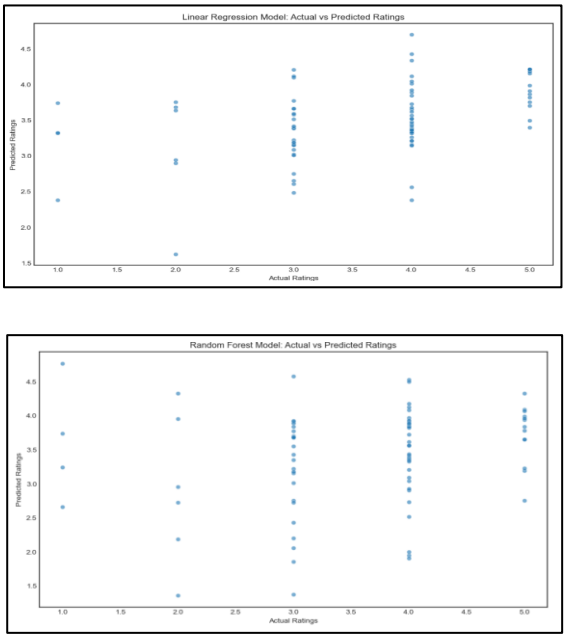

Fig.6&7. Linear Regression model: Actual vs. Predicated ratings and Random Forest model: Actual vs. Predicate ratings of Netflix-style user-movie interaction dataset.

V. CONCLUSION

This research compared movie rating prediction algorithms across two datasets: IMDb metadata-rich content and Netflix-style user-movie interactions. The results highlight the significant influence of dataset properties on model selection. On the IMDb dataset, gradient boosting algorithms, particularly CatBoost, achieved superior performance (93.74% accuracy, R^2=0.86) by effectively modeling complex interactions between content metadata features. In contrast, on the Netflix dataset, simpler models like Linear Regression performed better (R^2=0.72), aligning with the tendency of simple models to generalize better on sparse user-item interaction data.

Our findings emphasize the importance of tailored feature engineering strategies for different types of data. In content-based prediction, features like director reputation, genre, and cast were the most influential, while user demographic variables and temporal patterns were key for user-interaction predictions. This supports the notion that "historical values derived from past films exhibit potential to anticipate more accurate ratings" [20], but our study reveals that historical patterns are only effective within specific dataset contexts. The comparison highlights that the nature of the dataset determines whether a metadata-based or interaction-based model will perform better in real-world applications.

This research demonstrates that simple models excel with user interaction data, while ensemble methods perform better with rich metadata. These findings are valuable for streaming platforms to develop hybrid models that capitalize on both strengths. As content sites generate massive datasets, this research paves the way for more accurate and effective rating prediction systems tailored to various data environments. Future work should explore dynamic model selection and advanced network techniques to enhance prediction accuracy across diverse content ecosystems.

REFERENCES.

[1] W. R. Bristi, Z. Zaman and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944604.

[2] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 385-390, doi: 10.1109/ICSCCC.2018.8703320.

[3] Zahabiya Mhowwala, A. Razia Sulthana and Sujala D. Shetty, “Movie Rating Prediction using Ensemble Learning Algorithms” International Journal of Advanced Computer Science and Applications(IJACSA), 11(8), 2020.

[4] J. Bobadilla, F. Ortega, et al. (2013) "Recommender system survey" https://doi.org/10.1016/j.knosys.2013.03.012

[5] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[6] Wang, Huiqing & Zhao, Xiaofeng & Li, Yongjie & Xu, Jingwen & Liu, Shuai & Chen, Lianjie. (2024). Feature Engineering Strategies for Enhancing Movie Rating Prediction Models. 10.13140/RG.2.2.30872.20482.

[7] Kose, Alper, Can Kanbak, and Noyan Evirgen. "Performance comparison of algorithms for movie rating estimation." 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017.

[8] Ayesha Siddique, M Kamran Abid, Muhammad Fuzail, and Naeem Aslam. “Movies Rating Prediction Using Supervised Machine Learning Techniques”. International Journal of Information Systems and Computer Technologies, vol. 3, no. 1, Jan. 2024, pp. 40-56, doi:10.58325/ijisct.003.01.0062.

[9] X. Li, H. Zhao, Z. Wang and Z. Yu, "Research on Movie Rating Prediction Algorithms," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), Xiamen, China, 2020, pp. 121-125, doi: 10.1109/ICBDA49040.2020.9101282.

[10] Abidi, S.M.R., Xu, Y., Ni, J. et al. Popularity prediction of movies: from statistical modeling to machine learning techniques. Multimed Tools Appl 79, 35583–35617 (2020).

[11] Gupta, V., Jain, N., Garg, H. et al. Predicting attributes based movie success through ensemble machine learning. Multimed Tools Appl 82, 9597–9626 (2023).

[12] M. Marović, M. Mihoković, M. Mikša, S. Pribil and A. Tus, "Automatic movie ratings prediction using machine learning," 2011 Proceedings of the 34th International Convention MIPRO, Opatija, 2011, pp. 1640-1645.

[13] A. Kose, C. Kanbak and N. Evirgen, "Performance Comparison of Algorithms for Movie Rating Estimation," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 955-959, doi: 10.1109/ICMLA.2017.00-30.

[14] Kim, H., & Moon, J. (2024). Performance Comparison and SHAP Interpretation of Movie Box Office Prediction Models Based on CatBoost and PyCaret. Journal of Internet of Things and Convergence , 10 (5), 213–226.

[15] Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," in Computer, vol. 42, no. 8, pp. 30-37, Aug. 2009, doi: 10.1109/MC.2009.263.

[16] Lee, K., Park, J., Kim, I. et al. Predicting movie success with machine learning techniques: ways to improve accuracy. Inf Syst Front 20, 577–588 (2018).

[17] K. Pradeep, C. R. TintuRosmin, S. S. Durom and G. S. Anisha, "Decision Tree Algorithms for Accurate Prediction of Movie Rating," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 853-858, doi: 10.1109/ICCMC48092.2020.ICCMC-000158.

[18] Kharb, Latika & Chahal, Deepak & Vagisha,. (2020). Forecasting Movie Rating Through Data Analytics. 10.1007/978-981-15-5830-6_21.

[19] Abarja, Rudy Aditya, and Antoni Wibowo. "Movie rating prediction using convolutional neural network based on historical values." *Int. J* 8 (2020): 2156-2164.

[20] N. Darapaneni et al., "Movie Success Prediction Using ML," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2020, pp. 0869-0874, doi: 10.1109/UEMCON51285.2020.9298145.