Bangla Movies Popularity Prediction*

1st Ayesha Afroza Mohsin
Comp. Sci. and Engineering Dept.
Islamic University of Technology
Dhaka, Bangladesh
ayeshaafroza@iut-dhaka.edu

1st Lomatul Mahzabin
Comp. Sci. and Engineering Dept.
Islamic University of Technology
Dhaka, Bangladesh
mahzabin@iut-dhaka.edu

1st Shaira Sadia Karim
Comp. Sci. and Engineering Dept.
Islamic University of Technology
Dhaka, Bangladesh
shairasadia@iut-dhaka.edu

***Abstract*—This machine learning project aims to predict popularity of Bangla movies based on various features such as storyline, genre, director, actors, actresses, release date, runtime, average IMDb rating, trailer likes, views and comments etc. The dataset comprises approximately 200 Bangla movies sourced from platforms like Kaggle, IMDb, YouTube, Chorki, Wikipedia, TMDB, Bioscope and other online platforms and websites. Through machine learning models and data analysis, we seek to create a model that will provide insights into audience preferences, trends and factors influencing the success of a Bangla movie, and ultimately aid filmmakers, producers, and stakeholders making informed decisions about their cinematic works and thus contribute to the creative and financial success of the Bangla film industry.**

Github Link of the Project

I. Introduction

Amidst heightened competition in the film industry, there is an increasing need for effective tools that can forecast the success of movies.

Despite the ever-changing nature of the industry, the scarcity of comprehensive Bangla movie datasets, unlike in other industries, adds an additional layer of complexity to our predictive efforts. The limited availability of structured data poses a unique challenge, compelling us to draw insights from diverse sources and platforms.

Our dataset, comprising approximately 200 Bangla movies, sourced from diverse platforms collectively form the basis for building a predictive model that can unravel the intricate web of factors influencing the success of a movie.

Again, the Bangla film industry is at a critical juncture, experiencing a decline in the yearly production rate and a diminishing appeal, signifying a crucial moment for its future. Notably, only a very few Bangladeshi movies have achieved the status of being super hits, blockbusters, or all-time blockbusters. However, it is precisely during such challenging periods that innovative tools, like the predictive model we aim to develop, can offer valuable insights and strategies about movie production, marketing strategies, script selection, casting, and overall investment to rejuvenate the industry.

Moreover, the project aims to ensure that marketing efforts are targeted toward the aspects that most appeal to the target demographic, thus increasing overall audience engagement and a more satisfying viewer experience.

Through the exploration of our dataset and the development of a robust predictive model, we aim to unlock new possibilities for success in the dynamic landscape of Bangla cinema.

II. Background Study/Related Works

Traditionally, movie success has been measured by gross box office revenue [4]. However, there is a growing acknowledgment of additional significant parameters beyond revenue, such as positive audience response, which may not always align with financial gains.

Researchers in [5] forecasted IMDb ratings, while others in [1] categorized films according to budget, and another study by [2] classified Bollywood movies as either hits or flops by considering factors such as music and IMDb ratings. Other studies explored diverse factors like budget, IMDb votes, and screen count [3].

The incorporation of various elements like actors, directors, budget, and release details was investigated by researchers [3]. This corresponds to the concept that success is influenced by a combination of factors, as illustrated in the studies by [10], where factors such as 'who,' 'what,' and 'when' were taken into account.

Social media feedback and user reviews on platforms like IMDb, Twitter, and YouTube are recognized as influential factors [7]. Sentiment analysis, opinion mining [9], and user-generated content contribute to understanding audience sentiments, as demonstrated by various studies [8]. However, user reviews might exhibit bias, especially when provided by devoted fans of a particular actor or actress, potentially lacking impartiality. Again, most of them do not take critic reviews into account.

Challenges in data collection are evident across various studies. Researchers often tailor datasets based on their research requirements [6][8]. This diversity in data sources emphasizes the need for a comprehensive dataset for accurate analysis.

Despite extensive research, there is a notable gap in incorporating machine learning models into the Bangla film industry. Limited datasets specific to Bangla movies pose unique challenges.

The exploration of diverse features, challenges in data collection, and the unique characteristics of Bangla movies set the stage for our research. Our project aims to address these challenges by leveraging predictive models, including

Decision Tree, Random Forest, Linear Regression, and Support Vector Machine (SVM), for forecasting the success of Bangla movies.

III. Proposed Methodology/Approach

A. Architecture

- **Machine Learning Framework: scikit-learn**
- **Model Selection:**
  - **Decision Tree**
  - **Random Forest**
  - **Linear Regression**
  - **Support Vector Machine (SVM)**

  These models were selected based on their suitability for regression tasks. Decision Tree and Random Forest provide non-linear relationships, Linear Regression captures linear dependencies, and SVM is effective for complex, non-linear relationships.
- **Model Training and Validation:**
  We separate the data into two parts: one for training and another for testing, with an 80:20 split. Additionally, a cross-validation approach with 5 folds is employed to robustly assess the model's performance.
  Predictions are made using the test dataset, and various evaluation metrics, including Mean Absolute Error (MAE) and Accuracy, are utilized to measure the effectiveness of the model.
- **Evaluation Metrics:**
  - **Mean Absolute Error (MAE):** MAE provides insight into the average prediction error, measuring the absolute difference between predicted and actual values.
  - **Accuracy:** Accuracy offers a percentage-based assessment, providing a holistic view of model performance.
  - **Cross-Validation:** A cross-validation approach with 5 folds is employed to robustly assess the performance of the model and ensure its generalizability.
  - **Effect of Model Parameters:** The impact of different model parameters is analyzed to understand how variations influence model performance. This includes exploring hyperparameter tuning through techniques like **grid search**.
  - **Feature Importance Analysis:** An analysis of feature importance is conducted to identify the contribution of each feature to the predictions of the model. This helps identify which features significantly contribute to the model's decision-making process.

B. Dataset source

- **IMDb (Internet Movie Database):** Primary and comprehensive online database providing movie details, user ratings, and more.
- **Chorki:** Bangla streaming platform for additional localized insights.
- **Wikipedia:** General movie information and contextual details.
- **TMDB (The Movie Database):** A supplementary database offering diverse movie-related information.
- **Kaggle:** An additional source for datasets and insights.
- **YouTube API:** Utilized to extract movie trailer-related metrics, including views, likes, comments, and published dates, directly from YouTube.
- **External Websites:** Various reputable external websites were cross-referenced to supplement IMDb data and fill any gaps. These sources contribute to a more comprehensive dataset by providing additional details on cast, runtime, genre and related information.

C. Dataset description

The dataset used in this work is formed by 210 rows (movies) and 16 columns.

- **movieId:** A unique identifier for each movie.
- **title:** The title of the movie.
- **storyline:** A brief description or summary of the movie plot.
- **trailerLink:** Link to the movie trailer.
- **source-trailer:** Source of the trailer.
- **genre:** The genre or genres to which the movie belongs.
- **director:** The director of the movie.
- **starring:** The main cast or stars of the movie.
- **released year:** The year in which the movie was released.
- **runtime:** The duration of the movie in minutes.
- **IMDb avg rating:** The average rating of the movie in IMDb.
- **no of users (who rated):** The number of users who have rated the movie on IMDb.
- **views:** The number of views for the movie trailer.
- **likes:** The number of likes for the movie trailer.
- **comments:** The number of comments for the movie trailer.
- **trailerPublishedAt:** The date when the move trailer was published.

Fig. 1. Bangla movies dataset.

D. Correlation matrix

To gain insights into the relationships between numeric features in the movie content dataset, we employed a correlation matrix. The matrix provides a numerical representation of how pairs of numerical attributes are correlated. The correlation values range from -1, indicating a strong negative correlation, to 1, signifying a strong positive correlation. A correlation value of 0 denotes no correlation between the attributes.

From [Fig. 2], we can observe that strong positive correlation (0.94) between releasedYear and trailerPublishedAt indicates that trailers are typically published around the time a movie is released.

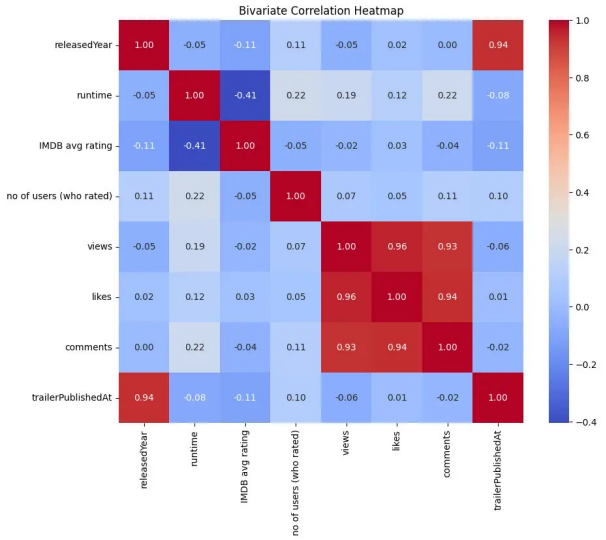

Fig. 2. Correlation matrix

Moderate negative correlation (-0.41) suggests that movies with shorter runtimes may have slightly higher IMDb ratings, and vice-versa.

High correlations (above 0.90) indicate a strong positive relationship among the number of views, likes, and comments on a trailer.

A weak positive correlation (0.11) exists between the year a movie was released and the number of users who rated it.

E. Visualizing the distribution of IMDb avg rating

The graph shows the Count vs IMDB Rating of the movies in our dataset. we can see that most of the movies are rated from 5 to 9 IMDB Score.

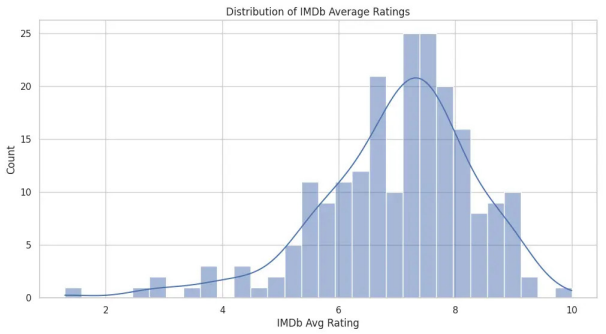

Fig. 3. Average Rating of the Movies

F. Visualizing the Pair Plotting

In order to visualize the affect of each feature column on our target variable, we plotted each feature against the IMDB avg rating and visualized the effect they have on the rating. It visualizes the relationships between pairs of variables, showcasing both the distributions of individual variables and the correlations between them.

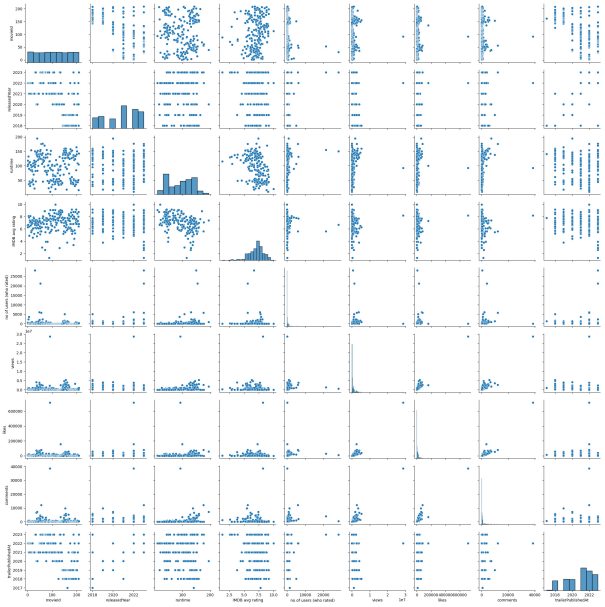

Fig. 4. Pair Plotting of the data

However, the use of pair plot for this specific dataset, especially with columns like movieId, title, storyline, trailerLink, etc., might not be the most appropriate. The reason is that pair plots are typically used for numerical variables, and some of the columns in the dataset, such as title, storyline, and trailerLink, are likely non-numeric. Afterwards, we decided to remove the following columns from the data-set: movieId, title, storyline, trailerLink, source_trailer, releasedYear, and trailerPushedAt. [Fig. 4]

G. Pre-processing steps

Several steps were taken to pre-process the data-set to make it easier for training machine learning models. Firstly we removed any leading or trailing spaces in the data. Some of the values in the data were multiple values that were comma separated or space separated and were thus converted into appropriate lists. The genre contained space separated values and the director and the starring features contained comma separated values. Further standardization was applied to have as few categories in the features as possible. For example we converted all 'Romantic' genre to 'Romance' genre. We also ensured all numeric data was being interpreted as such.

H. Feature engineering

The number of features present in the end dataset after multi-level encoding were 706 during the initial trial. After further contemplation we decided to remove all the categories that only had a single data point to represent it. We then Encoded it using multi-level binarizer and got 170 features instead.

I. Model implementation

The nature of our feature set was mostly categorical and the nature of our output variable (IMDb avg rating) was

continuous. We attempted to implement 4 models in total and compare the accuracy they displayed.

a) Decision Tree Regressor: Accuracy was 79.24%
b) Random Forest Regressor: Accuracy was 86.38%
c) Linear Regressor: Accuracy was 56.85%
d) Support Vector Regressor: Accuracy was 83.36%

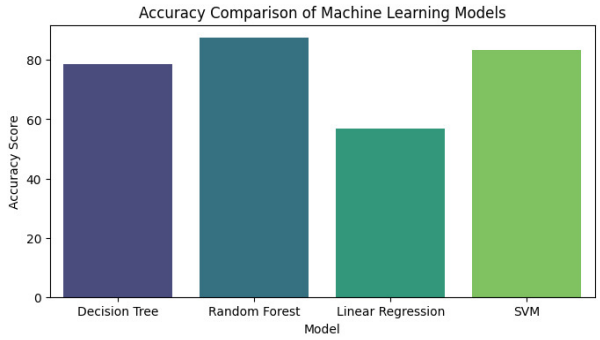

Fig. 5. Accuracy Comparison between Models

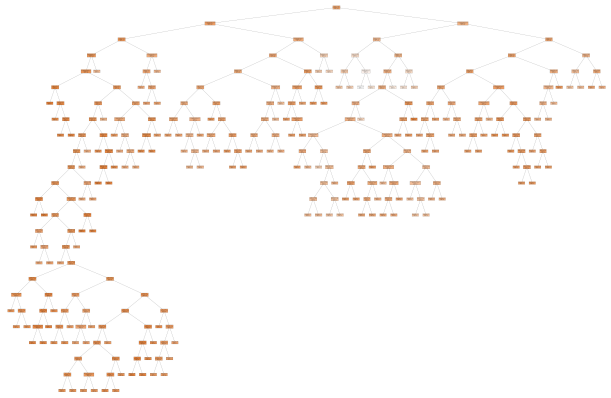

Fig. 6. Visualization of Decision Tree

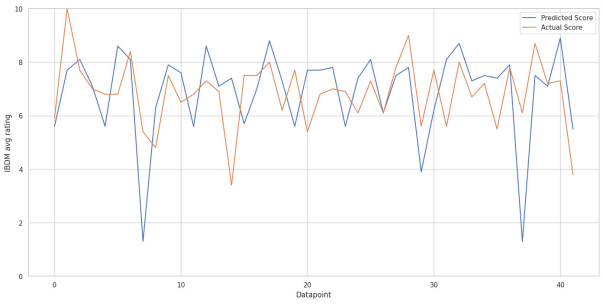

Fig. 7. Visualization of Decision Tree Regressor Predictions

IV. Results Analysis

We tried several metrics to assess the performance of the models we trained.

For Decision Tree model, we achieved a Mean Absolute Error of 1.24 degrees and an Accuracy of 79.24% on the test dataset. Cross-validation with k=5 resulted in MAE scores ranging from 0.96 to 1.48, with a mean MAE of 1.25. Through feature importance analysis, the model assigns higher importance to features like runtime and no of users (who rated), indicating that these factors play a significant role in predicting the IMDb average rating. Additionally, genres such as Action, Biography, Horror, Mystery, Social, Thriller, and War, as well as specific directors, contribute to the predictive power of the model. Through grid search, the best optimized model parameters were : max_depth: 30, min_samples_leaf: 4, min_samples_split: 15.

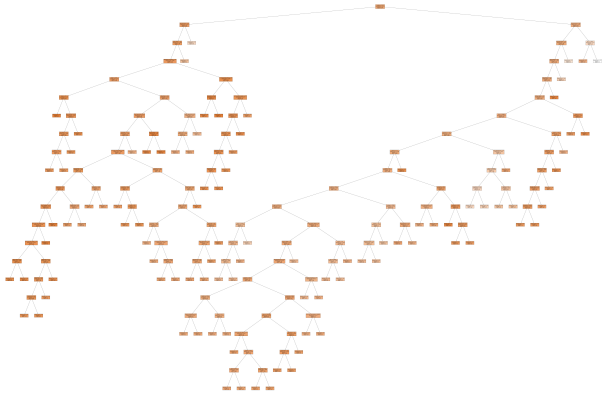

Fig. 8. Visualization of First Tree of Random Forest

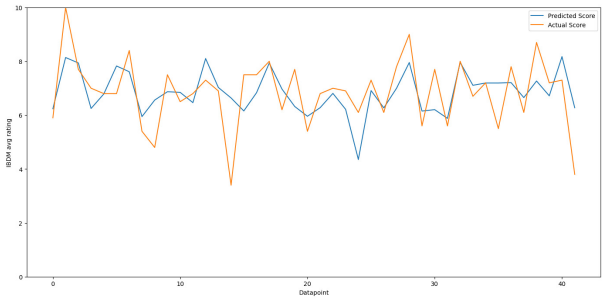

Fig. 9. Visualization of Random Forest Regressor Predictions

For Random Forest model, we achieved the best performance with the lowest Mean Absolute Error of 0.81 degrees and highest Accuracy of 86.38% on the test dataset. Cross-validation with k=5 resulted in MAE scores ranging from 0.79 to 1.14, with a mean MAE of 0.97, showing good generalization. Through feature importance analysis, the model assigns higher importance to features like the no of users (who rated) and views. Additionally, genres such as Drama and Action contribute significantly. Through grid search, the best optimized model parameters were : max_depth: 10, min_samples_leaf: 2, min_samples_split: 5, n_estimators: 100. So, increasing the number of trees and adjusting min_samples_leaf had notable effects on performance of Random Forest.

For Linear Regression model, we achieved a Mean Absolute Error of 2.67 degrees and an Accuracy of 56.85% on the test

dataset. Cross-validation with k=5 resulted in MAE scores ranging from 2.82 to 86.54, with a mean MAE of 22.91, thus indicating variability in performance across different subsets of the data, with one subset showing higher MAE. From feature importance analysis, we can see that the model is influenced by the no of users (who rated), genre and views.

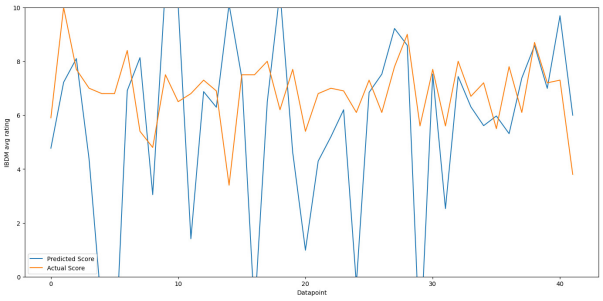

Fig. 10. Visualization of Linear Regression Regressor Predictions

For Support Vector Machine model, we achieved a Mean Absolute Error of 0.95 degrees and an Accuracy of 83.36% on the test dataset. Cross-validation with k=5 resulted in MAE scores ranging from 0.56 to 1.15, with a mean MAE of 1.01, thus indicating the consistent performance of the model across different data subsets, demonstrating strong generalization capabilities. So, the SVM Regressor demonstrates strong predictive performance with a low mean absolute error and high accuracy.

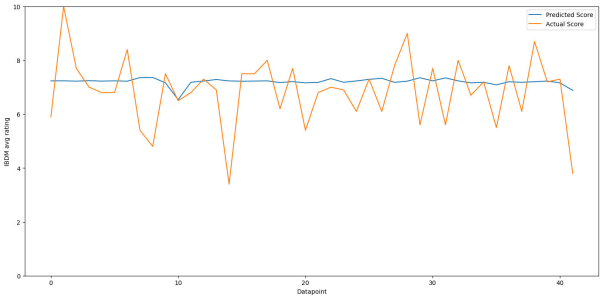

Fig. 11. Visualization of Support Vector Machine Regressor Predictions

V. Conclusion

As evident from the results, the Random Forest Regression yielded the highest accuracy, making it the preferred choice for the model. Random Forest is the ensemble version of Decision Trees so it's expected to perform better. The Linear Regression performed the worst. The Support Vector Regression performed relatively well in terms of accuracy not due to its ability to capture the relationships between the data points but due to tendency to output a relatively average value. The best model accuracy we could obtain was approximately 87.42%, as given by the Random Forest Regressor.

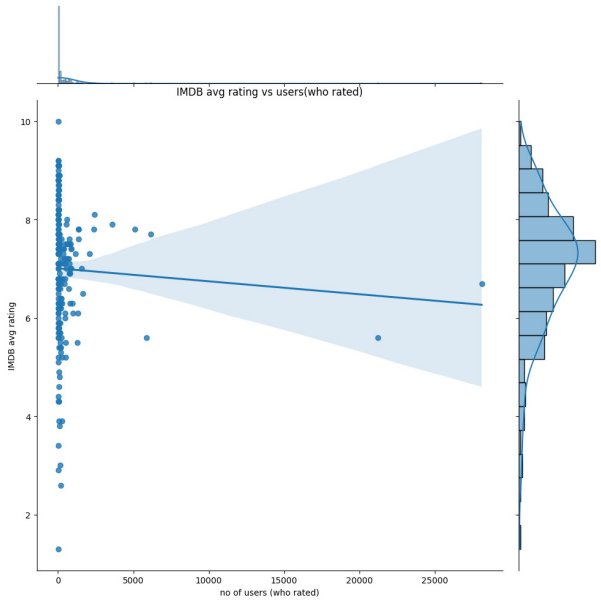

Fig. 12. Joint Plot IMDb avg rating vs no of users (who rated)

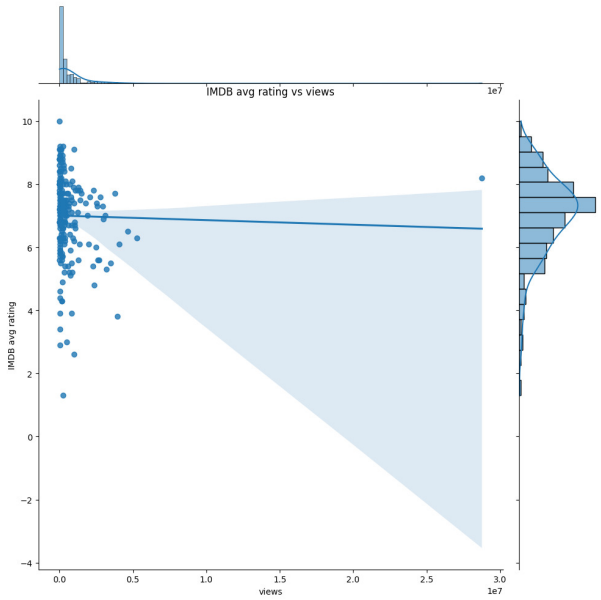

Fig. 13. Joint Plot IMDb avg rating vs views

References

[1] V. Subramaniyaswamy, V. V. M, V. P. R, and R. Logesh, "Predicting Movie Box Office Success using Multiple Regression and SVM," 2017 Int. Conf. Intell. Sustain. Syst., no. Iciss, pp. 182–186, 2017.

[2] G. Verma and H. Verma, "Predicting Bollywood Movies Success Using Machine Learning Technique," 2019 Amity Int. Conf. Artif. Intell., pp. 102–105, 2016.

[3] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, "Movie Success Prediction Using Data Mining," pp. 2015–2018, 2017.

[4] S. Gopinath, "Performance," no. May 2015, 2013.

[5] W. R. Bristi, "Predicting IMDb Rating of Movies by Machine Learning Techniques," 2019 10th Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, 2019.

[6] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," 2018 First Int. Conf. Secur. Cyber Comput. Commun., pp. 385–390, 2018.

[7] A. Bhave, "Role of Different Factors in Predicting Movie Success," vol. 00, no. c, 2015.

[8] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An

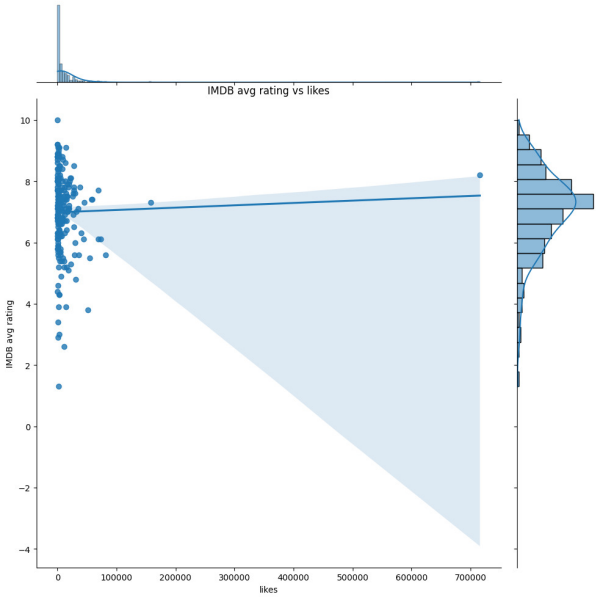

Fig. 14. Joint Plot IMDb avg rating vs likes

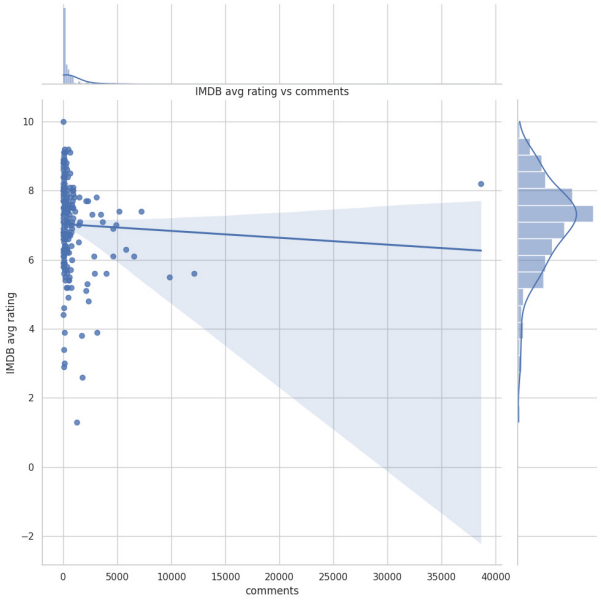

Fig. 15. Joint Plot IMDb avg rating vs comments

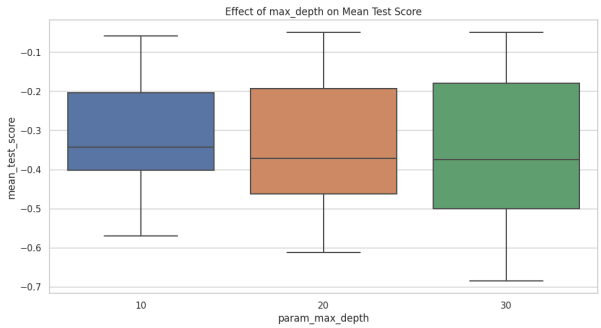

Fig. 16. Visualization of max depth on Mean test score of Decision Tree

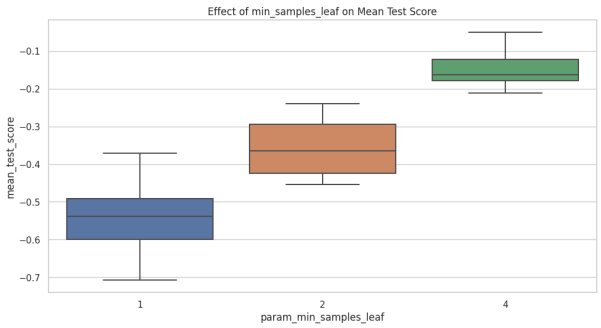

Fig. 17. Visualization of min samples leaf on Mean test score of Decision Tree

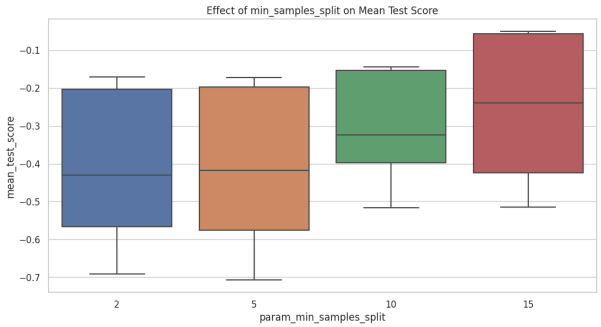

Fig. 18. Visualization of min samples split on Mean test score of Decision Tree

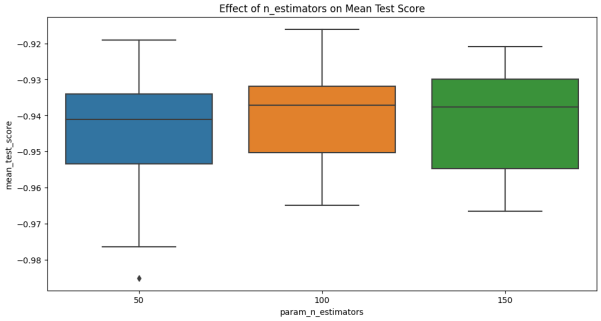

Fig. 19. Visualization of n estimators on Mean test score of Random Forest

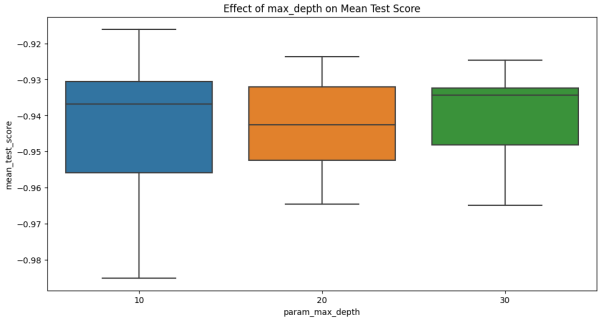

Fig. 20. Visualization of max depth on Mean test score of Random Forest

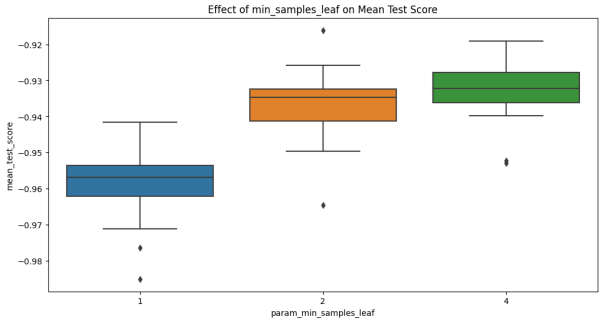

Fig. 21. Visualization of min samples leaf on Mean test score of Random Forest

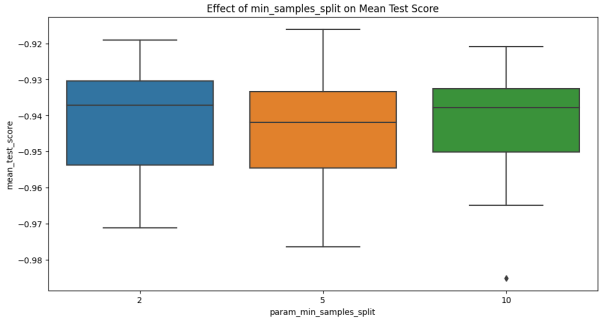

Fig. 22. Visualization of min samples split on Mean test score of Random Forest

Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction," pp. 933–937, 2015.

[9] A. Samad, H. Basari, B. Hussin, I. G. Pramudya, and J. Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization," Procedia Eng., vol. 53, pp. 453–462, 2013.

[10] M. Lash, S. Fu, S. Wang, and K. Zhao, "Early Prediction of Movie Success — What , Who , and When," vol. 1, pp. 345–349.