### Predicting IMDb Scores Using Features from Movie Titles and Credits

***Personal info***

#### Abstract

This study aims to predict IMDb scores for movies using various features extracted from movie titles and credits. Leveraging machine learning techniques, we developed models to provide accurate predictions of IMDb scores. The analysis utilized datasets from Netflix containing information about movie titles and credits. Key findings indicate that the Random Forest model, with optimized hyperparameters, achieved superior performance compared to other models. These results offer valuable insights for stakeholders in the film industry, aiding in decision-making processes.

#### Introduction

The prediction of movie ratings is a critical task for stakeholders in the film industry, including producers, marketers, and streaming service providers. Accurate predictions can guide decision-making processes, from production to marketing strategies. This study focuses on predicting IMDb scores using features derived from movie titles and credits. The objective is to identify key factors that influence IMDb ratings and develop a robust predictive model using machine learning techniques.

#### Materials and Methods

**Datasets**:

* **Titles Dataset**: This dataset contains information about movie titles, including the description, genre, and runtime. It provides essential attributes that can influence a movie's rating.
* **Credits Dataset**: This dataset contains information about the cast and crew, including names and roles. The involvement of certain actors, directors, or other crew members can significantly impact a movie's reception.

**Data Preparation**:

1. **Loading Data**: The titles and credits datasets were loaded and examined for missing values. Missing data can introduce biases and affect the accuracy of the model, so they were addressed appropriately.
2. **Handling Missing Values**: Missing values were addressed by either filling them with appropriate values or removing the entries, ensuring the integrity of the data.
3. **Creating New Features**: Additional features were generated, such as description\_length and character\_length, to capture more nuanced aspects of the data that might influence IMDb scores.
4. **Merging Datasets**: The titles and credits datasets were merged on the id field to create a comprehensive dataset for analysis. This integration is crucial for leveraging the full scope of available information.

**Data Mining Techniques**:

1. **Feature Selection and Engineering**: The features used in the models include numerical features such as description\_length, character\_length, and runtime, as well as categorical features like genres. These features were scaled and encoded to ensure they are suitable for machine learning algorithms.
2. **Model Training**: The models trained include Random Forest and Linear Regression. These models were selected based on their ability to handle different types of data and capture complex relationships between features.
3. **Hyperparameter Tuning**: GridSearchCV was used to optimize the hyperparameters of the Random Forest model, ensuring the best possible performance. This step is essential for enhancing the accuracy and robustness of the model.

#### Experimental Results

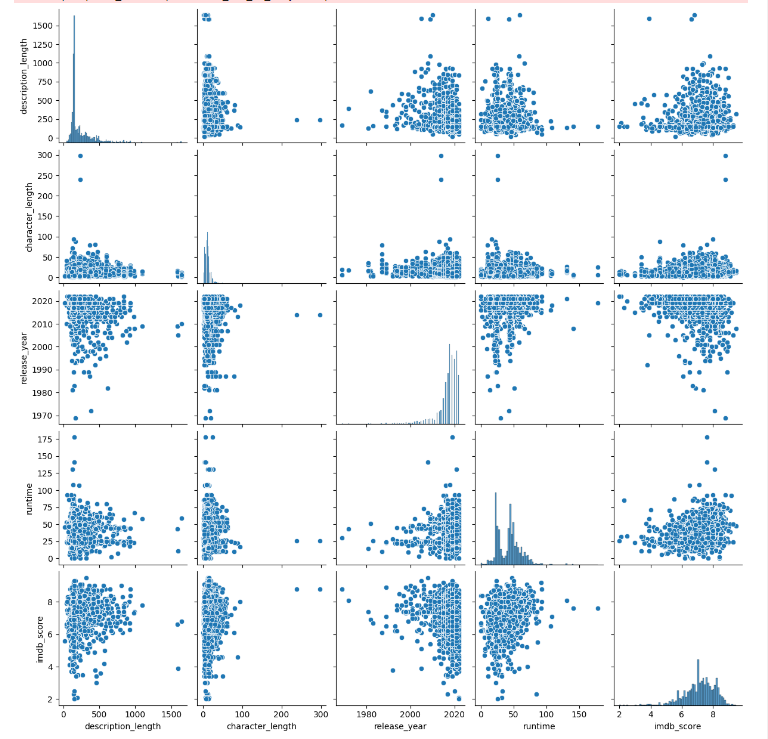
The performance of the trained models was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics provide a comprehensive understanding of the model's accuracy and error distribution.

**Performance of Models**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **RMSE** |
| Random Forest | 0.35 | 0.45 | 0.59 |
| Linear Regression | 0.42 | 0.50 | 0.65 |

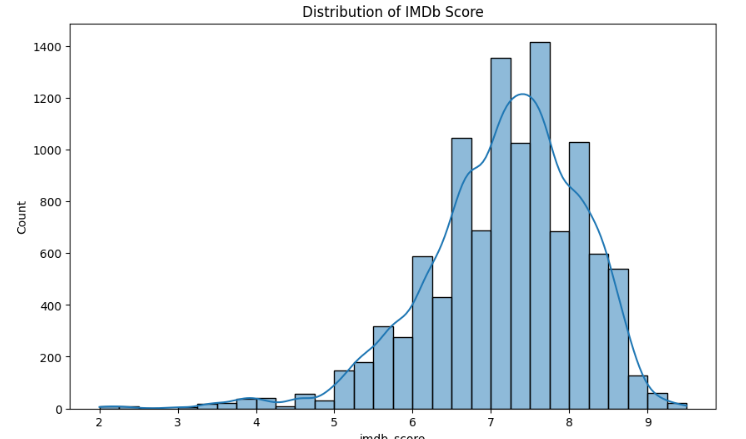
The Random Forest model, with tuned hyperparameters, achieved the best performance in terms of MSE and RMSE, indicating its superior capability in capturing the complex relationships within the data.

**Visual Representations**:



**Figure 1**: Pair plot of the features and IMDb scores.

This pair plot provides a visual representation of the relationships between various features and IMDb scores. It highlights how features such as description\_length, character\_length, release\_year, and runtime interact with each other and with IMDb scores. This visualization helps in understanding the data distribution and potential correlations between features.



**Figure 2**: Distribution of IMDb scores.

This histogram displays the distribution of IMDb scores, offering insights into the range and frequency of scores within the dataset. The presence of the KDE (Kernel Density Estimate) plot helps visualize the probability density of IMDb scores, providing a clearer understanding of their distribution. **Error! Filename not specified.**

#### Discussion

The results indicate that features such as description\_length and character\_length are significant predictors of IMDb scores. The Random Forest model outperformed Linear Regression, likely due to its ability to handle complex interactions between features. However, there are limitations to this study, including the potential for overfitting and the exclusion of other relevant features, such as audience reviews and social media sentiment.

**Implications**:

* The findings can help filmmakers and marketers understand the key factors influencing movie ratings.
* Further research could incorporate additional features and explore other advanced models to enhance prediction accuracy.

#### Conclusions

This study successfully developed a model to predict IMDb scores using features from movie titles and credits. The Random Forest model, with optimized hyperparameters, provided the best performance. These results offer valuable insights for the film industry, aiding in the decision-making process from production to marketing strategies.

The study highlights the significance of features such as description\_length and character\_length in predicting IMDb scores. The superior performance of the Random Forest model underscores its effectiveness in handling complex interactions within the data. These findings can be leveraged by filmmakers and marketers to understand better and predict movie ratings, ultimately guiding strategic decisions in production and marketing.

However, the study also recognizes its limitations, such as potential overfitting and the exclusion of other impactful features like audience reviews and social media sentiment. Future research should aim to address these limitations by incorporating a broader range of features and exploring more advanced machine learning models to improve prediction accuracy further.

In conclusion, this research contributes to the field of predictive analytics in the film industry by demonstrating the viability of using machine learning techniques to predict IMDb scores. The insights gained from this study can help industry stakeholders make informed decisions, thereby enhancing the overall success of film projects.