Film Forecast:
Predicting IMDb Ratings with Machine Learning

Jiyeon (Jenna) Woo, Mahnoor Shahid, and Kaylyn Nguyen

BAX 452: Machine Learning

Executive Summary

This project aimed to develop a predictive model for IMDb movie ratings using machine learning techniques. By analyzing a comprehensive dataset of movie characteristics, we developed and compared three advanced models: Random Forest, CatBoost, and LightGBM. Our analysis revealed key insights into factors influencing movie ratings and demonstrated the potential of machine learning in predicting audience reception.The key findings include the best-performing model for predicting movie success, which is CatBoost, achieving the lowest Root Mean Square Error (RMSE) of 0.8236 and the highest R^2 score of 0.5049. Key influential features include the number of voted users, genres, and movie duration. This model has potential applications in box office prediction, marketing strategies, and supporting production decisions.

Project Background and Motivation

The motivation for this project stemmed from the growing influence of data-driven decision-making in the entertainment industry. The movie industry relies heavily on audience perception and ratings for commercial success, and IMDb (Internet Movie Database) scores are one of the key rating systems used to determine reception of a movie. With over 500 movies being released in the United States each year, accurately predicting audience reception can provide valuable insights for filmmakers, studios, and marketers. Traditional methods of assessing a movie's potential success often rely on subjective evaluations or historical comparisons, which may not fully capture the complex interactions between various factors influencing viewer ratings. By leveraging machine learning, this project aims to develop a predictive model that can analyze key movie characteristics such as genre, duration, and audience engagement metrics to anticipate IMDb ratings with greater accuracy. This approach not only enhances understanding of what drives audience preferences but also supports more informed decision-making in casting, production, distribution, and promotional strategies.

Dataset

The analysis utilizes the Kaggle IMDb Movie Ratings Dataset, which includes key features such as director name, movie duration, lead actors, genres, audience engagement metrics language etc

Analyses

Methodology

Step 1. Import libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data analysis. Import libraries such as LightGBM, CatBoost, and sklearn for model building.

Step 2. Explore the Kaggle dataset and clean the data by dropping unnecessary columns, handling missing values appropriately, and encoding categorical variables. Split the cleaned data into training and test sets.

Step 3. Employ three advanced machine learning models, Random Forest, CatBoost, and LightGBM. Then choosing the most optimal one.

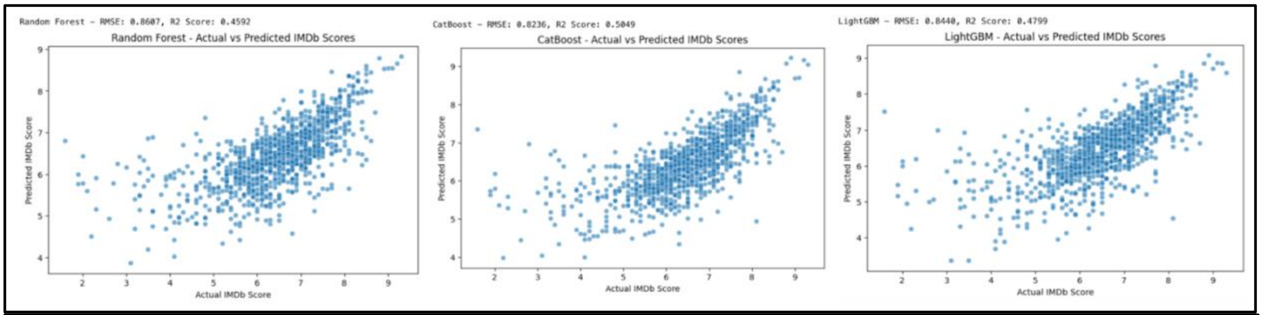
Step 4. Perform exploratory data analysis on the impact of certain features on the IMDb score.

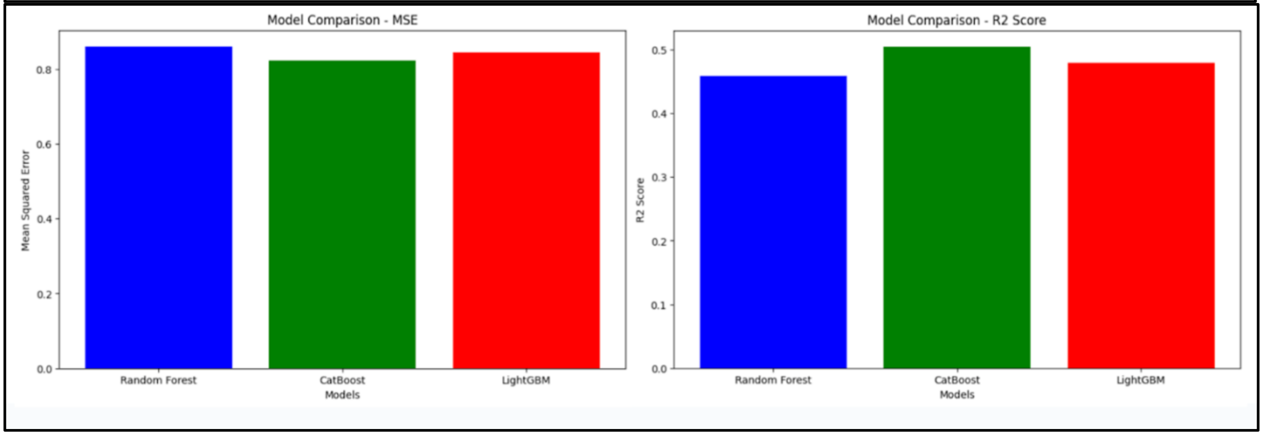
Step 5. Address the cold-start problem by accounting for pre-release factors (i.e. cast, director, and genre) and post-release factors (i.e. number of votes and reviews).

Step 6. Build IMDb Score Prediction model based on the most optimal model.

Model Performance Comparison

The predictive modeling methods used to assess IMDb scores for Random Forest, CatBoost, and LightGBM. The effectiveness is determined using Root Mean Square (RMSE) and R-squared (R^2) metrics.





The scatter plots and bar charts above compare and contrast the three different models. CatBoost outperforms both Random Forest and LightGBM in predicting IMDb ratings, achieving an RMSE of 0.8236 and an R^2 score of 0.5049, indicating it explains 50% of the variance. Its strengths include effective handling of categorical features, management of sparse data, built-in regularization to prevent overfitting, and the ability to capture complex feature interactions. Random Forest and LightGBM follow with RMSEs of 0.8607 and 0.8440, and R^2 scores of 0.4592 and 0.4799, respectively. Both models struggle with categorical data and capturing non-linear relationships, with Random Forest particularly overemphasizing numerical features. Overall, the results highlight CatBoost's superiority in structured data tasks with mixed feature types, consistent with industry findings.

EDA and Feature Importance Analysis

Next we look at the exploratory data analysis (EDA) and feature importance analysis of IMDb movie ratings, highlighting key trends and relationships within the dataset. Through various visualizations, we uncover insights into the distribution of scores, correlations with features, and the impact of genres and audience engagement on movie ratings. Figure 1 shows the distribution of IMDb scores, which is skewed slightly to the left. The most common IMDb scores are between 6 and 7, with very few movies being rated lower than 3 and higher than 8.

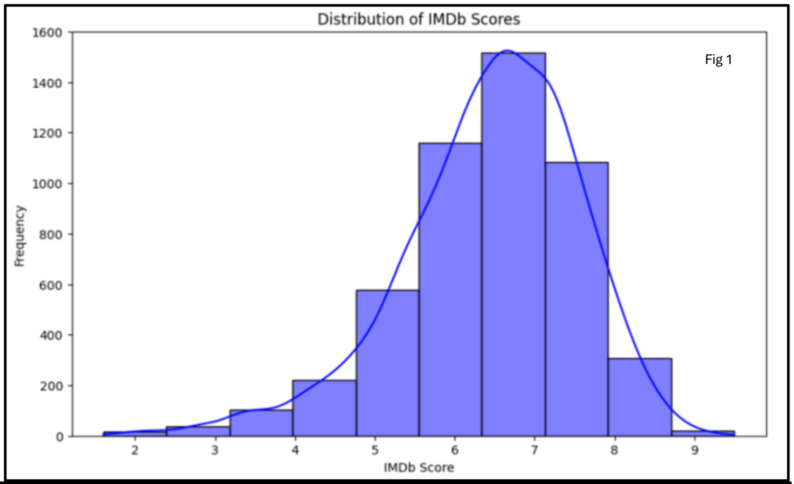


Fig 1

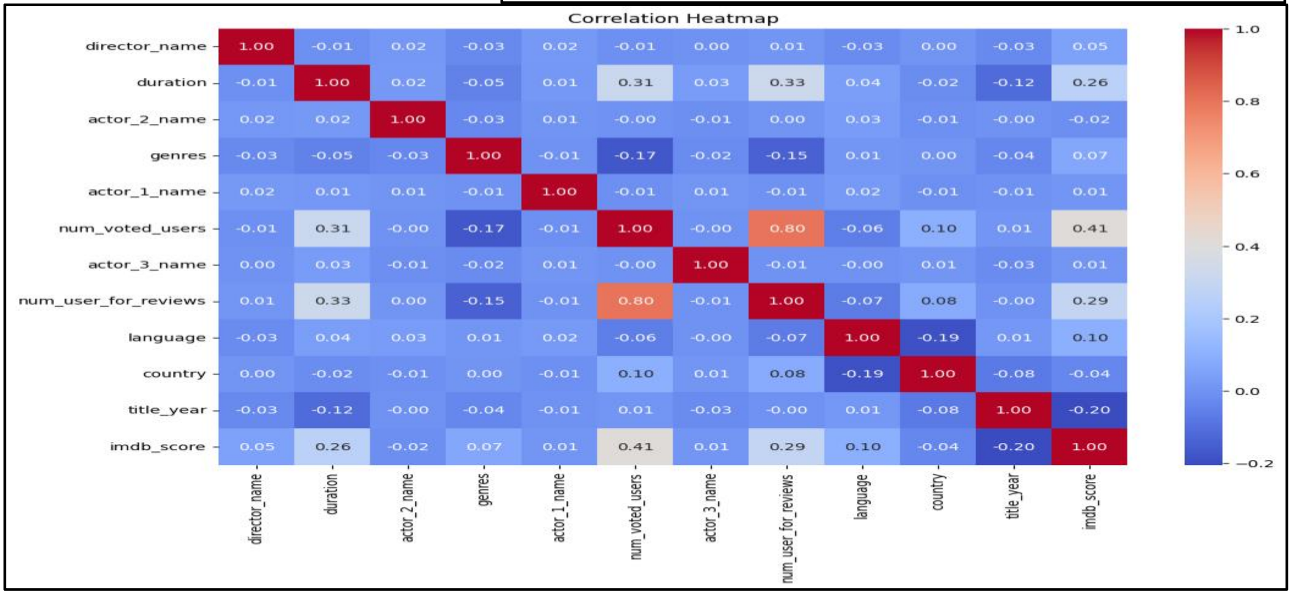


Fig 2

Our analysis also shows correlation between the features and the IMDb score as shown in fig 2 above.The most important features may be release year, number of votes, and number of reviews, but overall the correlation between most features and IMDb score is weak. This suggests that predicting movie ratings is complex and requires sophisticated modeling techniques.

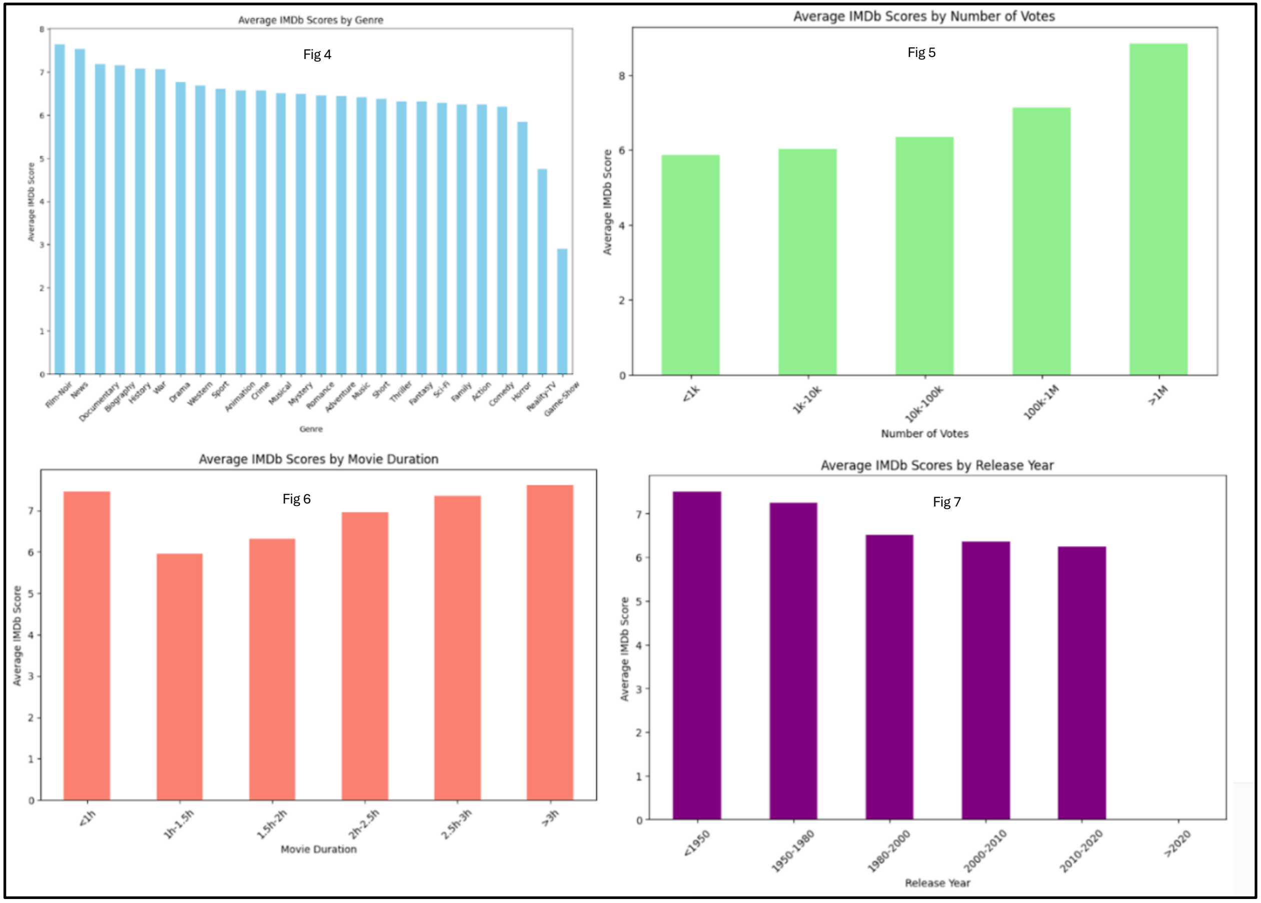
Regarding the range of scores over time, the range of scores grows wider while the median score follows a slightly downward trend The number of outlier values also increases over time, suggesting that more recent movies have a more polarizing reception (see figure 3, appendix).

Figure 4 below shows that the three genres with the highest average IMDb scores are film noir, news, and documentary. The three genres with the lowest average IMDb scores are horror, reality TV, and game shows. The trend suggests that informational genres tend to perform better in terms of ratings.

Figure 5 shows that the number of votes and average IMDb scores have a positive relationship. This trend suggests that widely watched and discussed films tend to be better received, possibly due to higher production quality or stronger audience engagement.

Figure 6 reveals that movies shorter than one hour and exceeding three hours tend to have the highest average ratings. This suggests that longer movies may be associated with higher quality content or stronger engagement, while shorter films could benefit from niche appeal.

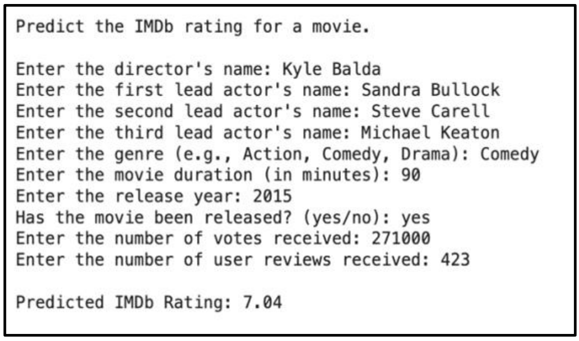
Figure 7 shows movies released before 1980 have the highest average ratings, while more recent movies tend to have lower average scores. This may indicate a difference in storytelling quality, or changes in audience expectations over time.



Moreover, feature importance analysis evaluates the significance of various features across three models, revealing consistent top features that are crucial for predicting ratings. The number of voted users emerges as the most important feature. Other significant features include genres, movie duration, and release year. Each model has unique emphases: Random Forest and CatBoost both place the most importance on number of user votes, duration, and genres while the LightGBM model put the most importance on genre, number of votes, and actor names (see figure 8, appendix)

IMDb Score Prediction Model

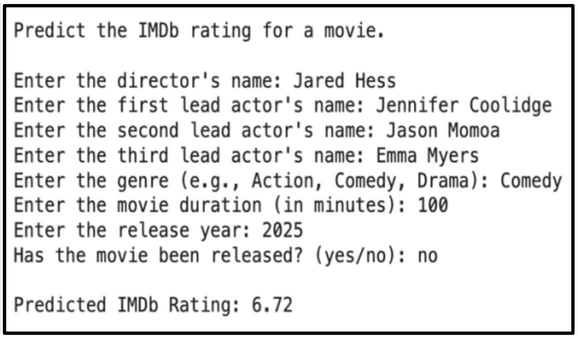
Figure 9 below shows the predicted IMDb ratings for two movies: "Minions," directed by Kyle Balda and released in 2015, got a rating of 7.04 based on its cast and user data, including 271,000 votes and 423 reviews. In contrast, "The Minecraft Movie," directed by Jared Hess and set for release in 2025, has a predicted IMDb rating of 6.72, despite its cast including Jennifer Coolidge and Jason Momoa



```
Predict the IMDb rating for a movie.

Enter the director's name: Kyle Balda
Enter the first lead actor's name: Sandra Bullock
Enter the second lead actor's name: Steve Carell
Enter the third lead actor's name: Michael Keaton
Enter the genre (e.g., Action, Comedy, Drama): Comedy
Enter the movie duration (in minutes): 90
Enter the release year: 2015
Has the movie been released? (yes/no): yes
Enter the number of votes received: 271000
Enter the number of user reviews received: 423

Predicted IMDb Rating: 7.04
```



```
Predict the IMDb rating for a movie.

Enter the director's name: Jared Hess
Enter the first lead actor's name: Jennifer Coolidge
Enter the second lead actor's name: Jason Momoa
Enter the third lead actor's name: Emma Myers
Enter the genre (e.g., Action, Comedy, Drama): Comedy
Enter the movie duration (in minutes): 100
Enter the release year: 2025
Has the movie been released? (yes/no): no

Predicted IMDb Rating: 6.72
```

.

The cold-start problem presents a significant challenge in predicting audience engagement for new or upcoming movies, as they often lack sufficient metrics such as votes and reviews. To tackle this issue, we propose a two-model approach that enhances prediction accuracy. The pre-release model leverages key movie characteristics, such as the director, cast, genre to forecast IMDb ratings before a film's release. Once the movie is out, the post-release model takes over, integrating audience metrics like the number of votes and reviews to refine these predictions. This distinction is crucial, for instance, consider two films with identical pre-release predictions but vastly different audience engagement: Movie A, with 500,000 votes and 30,000 reviews, likely has a more reliable rating than Movie B, which has only 5,000 votes and 500 reviews. By incorporating audience feedback, the post-release model can significantly enhance prediction accuracy, aligning more closely with actual IMDb ratings. Additionally, feature engineering will use historical averages for directors, actors, and genres, providing meaningful initial estimates for new films and further improving the predictive framework.

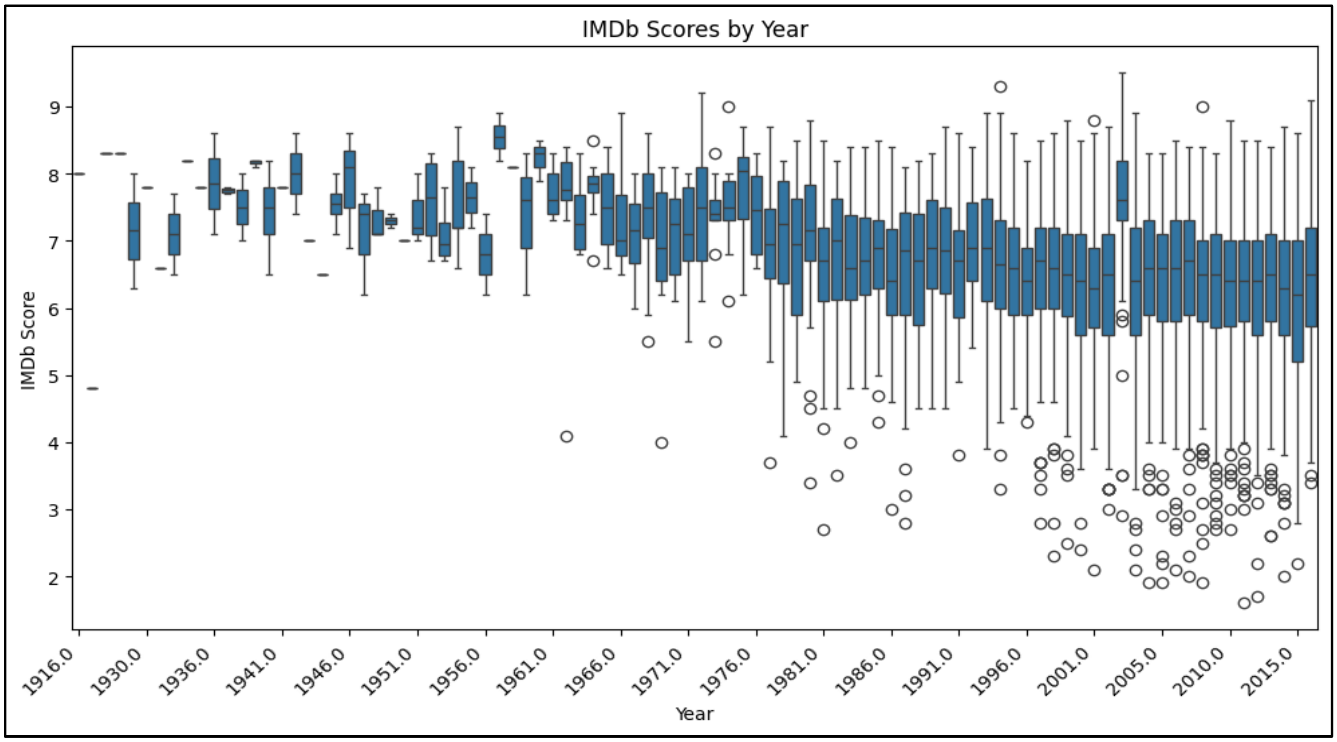
Recommendations

For IMDb score predictions, it is recommended to use CatBoost due to its lowest RMSE and highest R² score, as well as its ability to handle complex feature interactions. The focus should be on feature engineering, emphasizing audience engagement metrics, genre classification, director and actor historical performance, and temporal trends in movie ratings. A two-model approach is suggested, developing separate models for pre-release predictions based on movie characteristics and post-release refinement that incorporates audience metrics. Additionally, it may be beneficial to create genre-specific prediction models to account for significant rating variations across different movie types. Prioritizing features related to user votes and reviews is essential, along with developing strategies to address challenges posed by movies with limited audience interaction.

Conclusion

Our machine learning approach showcases the potential of data-driven insights in understanding movie ratings, utilizing advanced models and comprehensive feature analysis to offer valuable tools for the film industry. The analysis highlights the complex nature of movie ratings, influenced by factors such as audience engagement, genre characteristics, movie duration, and temporal trends. CatBoost emerged as the best-performing model. This model can inform box office predictions, marketing strategies, and production decisions. We also addressed the cold-start problem by developing separate pre-release and post-release models and incorporating feature engineering with historical averages. Future work could focus on advanced features, ensemble techniques, and genre-specific models, reinforcing the power of machine learning in delivering insights into audience perception and movie quality.

Appendix



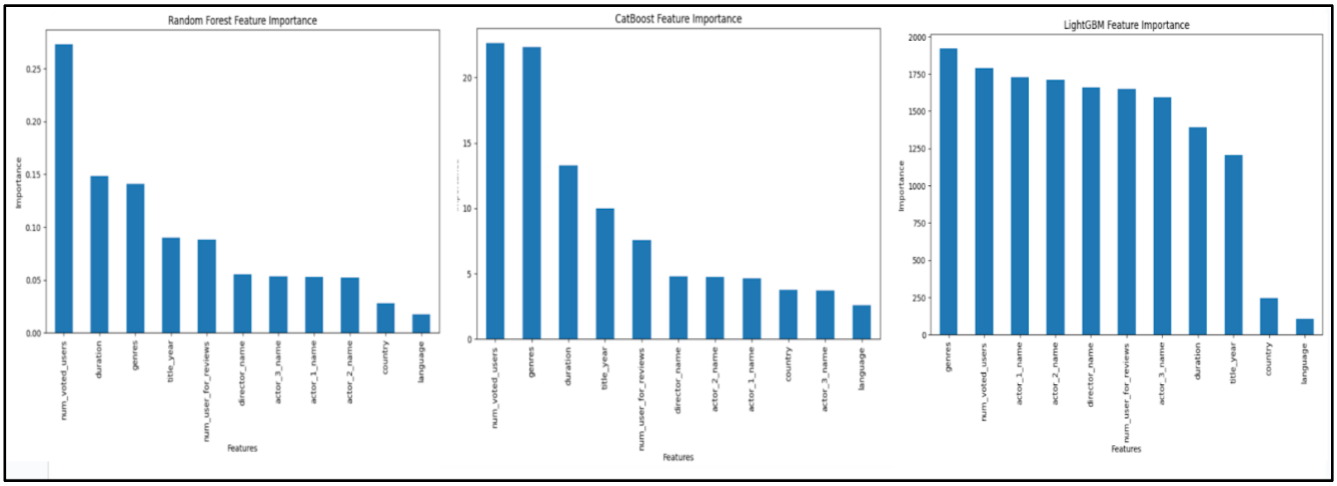


Fig 8