Predicting and Analyzing IMDb Ratings of Indian Movies Using Machine Learning

Results and Analysis

Among the evaluated models, ensemble-based approaches such as Random Forest and Gradient Boosting consistently outperformed linear models. These models were better able to capture non-linear relationships between features and IMDb ratings.

Key findings include:

- **Number of votes** is one of the strongest predictors of IMDb rating, reflecting audience engagement.
- **Contributor history** (actors and directors with consistently higher past ratings) positively influences predicted ratings.
- **Genre composition** plays a measurable role, with certain genres showing higher average ratings.
- Temporal trends indicate gradual changes in rating distributions over time.

	feature	importance
17	main_genre_Documentary	0.105287
22	main_genre_Horror	0.092705
10	decade	0.075068
3	cast_prior_avg	0.073900
11	main_genre_Action	0.050133
18	main_genre_Drama	0.049300
19	main_genre_Family	0.044370
24	main_genre_Musical	0.043995
14	main_genre_Biography	0.040331
5	dir_prior_avg	0.039963
0	votes	0.037222
20	main_genre_Fantasy	0.030785
16	main_genre_Crime	0.028539
8	movie_deg_max	0.025840
12	main_genre_Adventure	0.024840
7	movie_deg_mean	0.024306
26	main_genre_Romance	0.023119
9	movie_pr_mean	0.022865
4	cast_max_prior_count	0.020935
29	main_genre_Thriller	0.020227

Figure 1: Model performance comparison across evaluation metrics

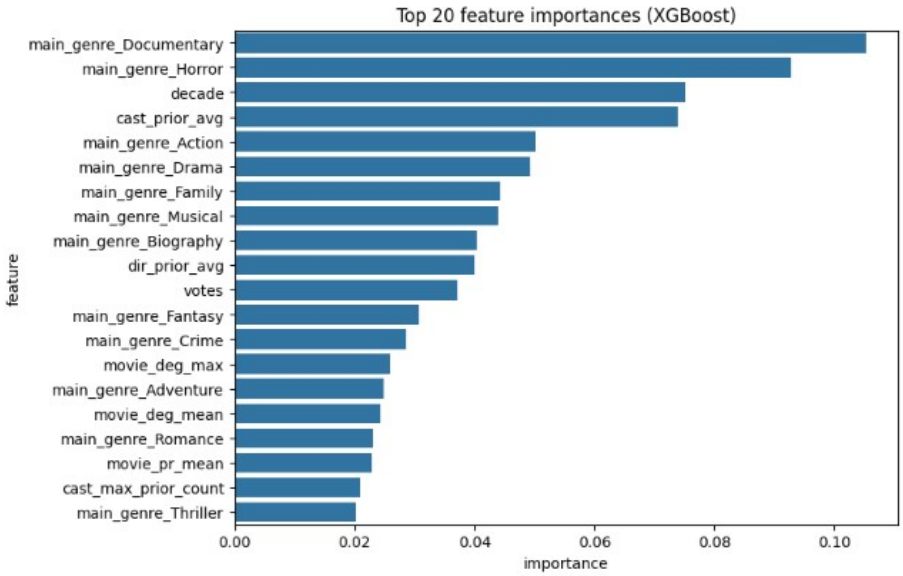

Figure 2: Feature importance visualization for the final model

Conclusion

This project demonstrates the effectiveness of machine learning techniques in analyzing and predicting IMDb ratings for Indian movies. Through systematic data cleaning, feature engineering, and model evaluation, meaningful insights were derived from a complex real-world dataset.

The final model achieved strong predictive performance and highlighted key factors influencing movie ratings, particularly audience engagement and contributor consistency. While alternative approaches such as deep learning or natural language processing on reviews were considered, they were not pursued due to data and scope constraints.

Overall, this project reflects the core learning objectives of the course, including iterative model development, feature engineering, and critical evaluation of machine learning results. Future work could explore incorporating textual reviews, social media signals, or more advanced representation learning techniques to further enhance predictive accuracy.

References

- IMDb Dataset – Kaggle
- scikit-learn Documentation
- Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*