EDA and predictive modeling of IMDB movie ratings

Insights and Conclusions

Insights

Model Performance:
The Linear Regression model achieved an R² score of 0.30 and a Mean Squared Error (MSE) of 0.50. This indicates that the model explains 30% of the variance in IMDb ratings, suggesting that there are other factors not captured by the selected features. The (MSE) indicates that the predictions are moderately close to the actual values, but there is still substantial room for improvement.

Data Cleaning: The preprocessing steps, including handling missing data and encoding categorical variables (like Director), were crucial in ensuring that the dataset was ready for modeling.

The analysis showed that Revenue (Millions), Runtime (Minutes), Votes, and Director had varying degrees of influence on the IMDb ratings, with certain patterns emerging, but these relationships are not purely linear and may require more complex models to be fully understood.

There is a weak positive correlation between a movie's Revenue and its IMDB Rating. Many high-revenue films have lower ratings than expected This suggests that while box-office success is an indicator of popularity, it does not necessarily mean a higher-quality film in terms of audience satisfaction.

The distribution of IMDB ratings shows a tendency towards higher ratings, with the mean rating being above average

Next Steps

Moving beyond linear regression to more complex models, such as Random Forest or XGBoost, could improve prediction accuracy. These models handle non-linearity better and are more adept at capturing intricate relationships between features.

Including additional features like genre, release year, budget, or cast could further improve model performance and provide a more comprehensive picture of factors influencing movie ratings.

corporating more diverse and external data sources, such as critic reviews, social media sentiment, or audience demographics, could provide a richer feature set for more accurate predictions.