IMDB Movie Rating Prediction

Project Final Report

Submitted by:
Yamineesh Kanaparthy
yamineesh@gmail.com

Table of Contents

INTRODUCTION

In the rapidly evolving landscape of the entertainment industry, movies play a pivotal role in captivating audiences and shaping cultural narratives. With the advent of online platforms, such as IMDb, audiences have gained the power to influence the success of a movie through ratings and reviews. In this context, harnessing the capabilities of machine learning to predict movie ratings becomes a compelling avenue for filmmakers, producers, and stakeholders. This report delves into the process of developing a machine learning model to predict movie ratings on the IMDb platform.

Project Management Report

Business Understanding

In the contemporary entertainment industry, movies have transcended from mere sources of entertainment to cultural phenomena that impact society and drive economic growth. Filmmakers, production companies, and distributors are constantly striving to create movies that resonate with audiences and yield substantial returns. IMDb, as a prominent online platform for movie information and reviews, plays a pivotal role in this ecosystem.

For movie industry stakeholders, understanding the factors that contribute to a movie's success on platforms like IMDb is of paramount importance. Ratings and reviews on IMDb can significantly influence a movie's reach, profitability, and critical acclaim. This necessitates the need for predictive models that can provide insights into the potential IMDb rating of a movie before its release. Such models can aid in refining marketing strategies, optimizing production budgets, and enhancing overall decision-making processes in the industry.

Problem Statement

The problem at hand involves developing a machine learning model to predict the IMDb rating of movies prior to their release. This predictive model aims to leverage historical data from IMDb and related sources to accurately forecast a movie's likely reception on the platform. The primary objectives are as follows:

1. **Data Collection and Preparation:** Gather and preprocess a comprehensive dataset containing relevant features such as genre, cast, director, production budget, release date, and potentially social media buzz.
2. **Feature Selection and Engineering:** Identify and engineer key features that have a substantial impact on movie ratings. These features could include the historical performance of the director, popularity of cast members, genre trends, and more.
3. **Model Building:** Develop and fine-tune machine learning algorithms capable of capturing complex relationships between the selected features and IMDb ratings. Potential algorithms include regression models, ensemble methods, and neural networks.

4. **Model Evaluation:** Employ appropriate evaluation metrics to assess the model's predictive performance. Techniques like cross-validation and holdout testing will be used to ensure robustness.
5. **Insights Generation:** Interpret the model's output to gain insights into the factors driving IMDb ratings. These insights can inform decision-making processes in the movie industry.

By successfully addressing these objectives, this project aspires to provide a tool that assists movie industry stakeholders in making informed decisions, thereby enhancing the overall quality and impact of movies released on the IMDb platform.

Overarching Objectives

The overarching objectives of this project are to create a predictive machine learning model that accurately forecasts the IMDb ratings of movies before their release. By achieving this, the project aims to provide valuable insights to stakeholders in the movie industry, aiding them in making informed decisions about production budgets, marketing strategies, and overall film quality. The model's predictions can serve as a guide to optimizing resources, enhancing audience engagement, and increasing the likelihood of a movie's success on the IMDb platform.

Project Plan and Timelines

The project plan below lists the detailed tasks, dependencies, and the timeline schedule to deliver the project.

Task List	Completion Date
Business Understanding	10th Aug 2023
Resource Planning and Scheduling	10th Aug 2023
Data Understanding	10th Aug 2023
Define Problem Statement	10th Aug 2023
Data Initial Exploration	11th Aug 2023
Data Cleaning	11th Aug 2023
Exploratory Data Analysis	12th Aug 2023
Data Preparation	12th Aug 2023
Data Modelling	13th Aug 2023
Optimization and Testing	14th Aug 2023
Documenting the Results	15th Aug 2023
Final Report Submission	16th Aug 2023

Timelines of Project Tasks

Project Strategy

The project will be executed in a systematic manner, following a structured strategy:

1. **Data Collection and Preprocessing:** Acquire a diverse and comprehensive dataset containing relevant features from IMDb and external sources. Cleanse and preprocess the data to ensure its quality and suitability for analysis.
2. **Feature Selection and Engineering:** Conduct exploratory data analysis to identify influential features. Engineer new features if necessary and eliminate redundant ones to enhance model performance.
3. **Model Selection and Training:** Experiment with a range of machine learning algorithms, including linear regression, decision trees, random forests, and neural networks. Fine-tune hyperparameters and assess models' performance using cross-validation techniques.
4. **Model Evaluation:** Evaluate the models' predictive accuracy using appropriate evaluation metrics such as Root Mean Squared Error (RMSE). Employ holdout testing to validate models' generalization capabilities.
5. **Interpretation of Results:** Translate technical findings into actionable insights for industry stakeholders.
6. **Risk Mitigation:** Continuously monitor and address potential challenges, such as data quality issues, model overfitting, or lack of interpretability. Implement appropriate regularization techniques and ensure robustness of the final model.

Risks and Challenges

1. **Data Quality and Availability:** The project's success relies heavily on the quality and availability of data. Incomplete or biased data could lead to inaccurate predictions. Mitigation involves thorough data preprocessing, handling missing values, and potentially sourcing external data to enrich the dataset.

2. **Feature Engineering Complexity:** Identifying engineering features that have a strong impact on IMDb ratings might be challenging. Domain expertise and creative feature engineering techniques will be required to capture nuanced relationships.

3. **Model Overfitting:** Complex models might be overfit to noise in the data, resulting in poor generalization to unseen movies. Regularization techniques, cross-validation, and careful hyperparameter tuning will be employed to mitigate overfitting risks.

4. **Interpretability:** Some advanced machine learning algorithms, like deep neural networks, lack interpretability. Striking a balance between predictive power and interpretability will be essential to provide actionable insights to stakeholders.

5. **Changing Trends:** Movie industry trends are dynamic and can change rapidly due to various external factors. The model's effectiveness could be impacted by shifts in audience preferences, social trends, or unforeseen events.

6. **Ethical Considerations:** The model's predictions could influence production decisions and resource allocation. Ensuring that the model's recommendations are not biased or discriminatory is crucial to maintaining fairness and ethical standards.

By proactively addressing these risks and challenges throughout the project's lifecycle, the team aims to develop a robust and accurate predictive model that contributes meaningful insights to the movie industry.

Metadata

The provided Data Set from imdb is a 18.6 MB .csv file with 5,043 rows and 28 columns of data. The data set includes,

- Descriptive Metadata: Eg: Director Name, Actor Name, Facebook Likes etc.
- Structural Metadata: Eg: Genre, Color, Duration etc.
- Administrative Metadata: Eg: Content Rating, Title Year, Aspect Ratio.

Data Analysis Tools/ Software

For this project, we plan to use a combination of various tools to deliver the goals.

Data Visualization and Exploration- Python, Excel.
Data Preparation and Modelling- Python

Data Quality Issues

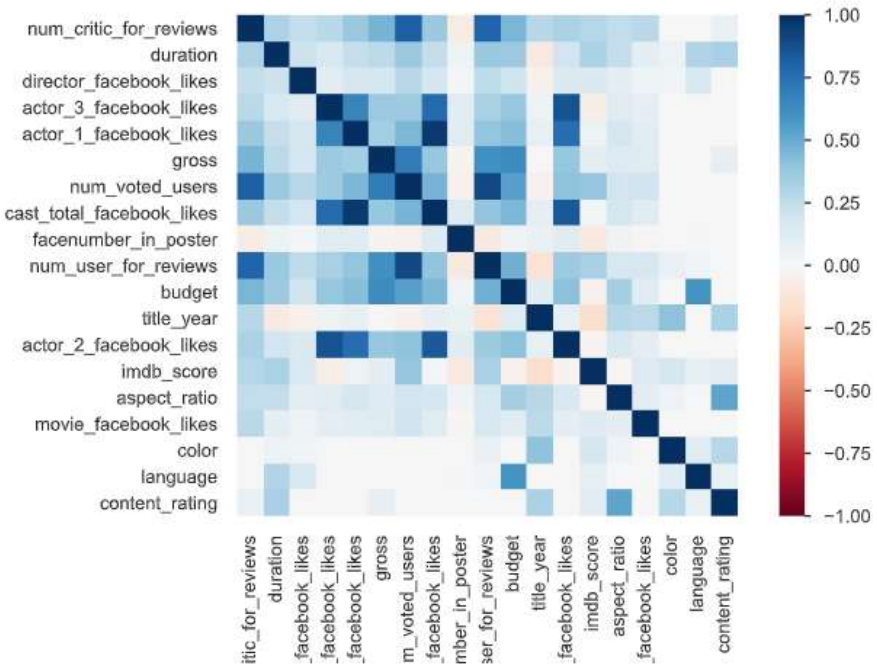
Below table lists the data quality issues,

Dataset has 45 (0.9%) duplicate rows	**Duplicates**
num_critic_for_reviews is highly overall correlated with num_voted_users and 1 other fields	**High correlation**
actor_3_facebook_likes is highly overall correlated with actor_1_facebook_likes and 2 other fields	**High correlation**
actor_1_facebook_likes is highly overall correlated with actor_3_facebook_likes and 2 other fields	**High correlation**
gross is highly overall correlated with num_voted_users and 2 other fields	**High correlation**

num_voted_users is highly overall correlated with num_critic_for_reviews and 3 other fields	**High correlation**
cast_total_facebook_likes is highly overall correlated with actor_3_facebook_likes and 2 other fields	**High correlation**
num_user_for_reviews is highly overall correlated with num_critic_for_reviews and 2 other fields	**High correlation**
budget is highly overall correlated with gross and 2 other fields	**High correlation**
actor_2_facebook_likes is highly overall correlated with actor_3_facebook_likes and 2 other fields	**High correlation**
aspect_ratio is highly overall correlated with content_rating	**High correlation**
language is highly overall correlated with budget	**High correlation**
content_rating is highly overall correlated with aspect_ratio	**High correlation**
color is highly imbalanced (75.0%)	**Imbalance**
language is highly imbalanced (88.8%)	**Imbalance**
content_rating is highly imbalanced (50.4%)	**Imbalance**
director_name has 104 (2.1%) missing values	**Missing**
director_facebook_likes has 104 (2.1%) missing values	**Missing**
gross has 884 (17.5%) missing values	**Missing**

plot_keywords has 153 (3.0%) missing values	**Missing**
content_rating has 303 (6.0%) missing values	**Missing**
budget has 492 (9.8%) missing values	**Missing**
title_year has 108 (2.1%) missing values	**Missing**
aspect_ratio has 329 (6.5%) missing values	**Missing**
budget is highly skewed (γ1 = 48.15743539)	**Skewed**
director_facebook_likes has 907 (18.0%) zeros	**Zeros**
actor_3_facebook_likes has 89 (1.8%) zeros	**Zeros**
facenumber_in_poster has 2152 (42.7%) zeros	**Zeros**
actor_2_facebook_likes has 55 (1.1%) zeros	**Zeros**
movie_facebook_likes has 2181 (43.2%) zeros	**Zeros**

Below is a correlation plot of the features. The plot highlights the highly correlated features in a darker shade.



Correlation Plot

Data Cleanup

To prepare the data for analysis, we performed a few initial steps.

- Observations with missing data were excluded from the analysis.
- 33 duplicate rows of data were excluded from the data set.
- Highly Correlated Feature pairs (Correlation Coefficient > 0.8) were identified and one of the pairs was dropped to avoid multicollinearity.

Data Initial Exploration and Overview

- Here is an overview of the dataset that was used for analysis.

Dataset statistics

Number of variables	28
Number of observations	5043
Missing cells	2698
Missing cells (%)	1.9%
Duplicate rows	45
Duplicate rows (%)	0.9%
Total size in memory	1.1 MiB
Average record size in memory	224.0 B

Variable types

Categorical	3
Text	9
Numeric	16

- The data set had 2097 lead actors or ‘actor_1’ in the data set.

actor_1_name
Text

Distinct	2097
Distinct (%)	41.6%
Missing	7
Missing (%)	0.1%
Memory size	39.5 KiB

- Comedy and drama were the most prominent genres.

genres
Text

Distinct	914
Distinct (%)	18.1%
Missing	0
Missing (%)	0.0%
Memory size	39.5 KiB

More details

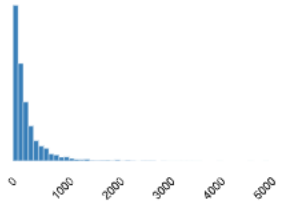
- An average of 273 user reviews were available for the movies in the data.

num_user_for_reviews

Real number (ℝ)

HIGH CORRELATION

Distinct	954	**Minimum**	1
Distinct (%)	19.0%	**Maximum**	5060
Missing	21	**Zeros**	0
Missing (%)	0.4%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	272.77081	**Memory size**	39.5 KiB



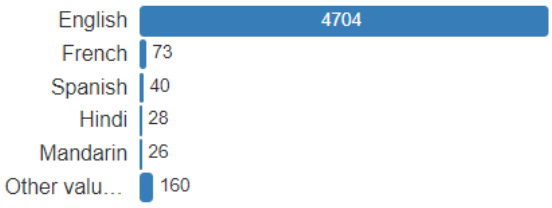
- Most of the reviews included were for English language movies.

language

Categorical

HIGH CORRELATION IMBALANCE

Distinct	47
Distinct (%)	0.9%
Missing	12
Missing (%)	0.2%
Memory size	39.5 KiB



- Movies produced from ‘USA’ were the major inclusions in the data.

country

Text

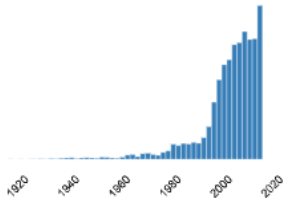
Distinct	65
Distinct (%)	1.3%
Missing	5
Missing (%)	0.1%
Memory size	39.5 KiB

- Movies released between 1916 and 2016 were included in the data.

title_year
Real number (ℝ)

MISSING

Distinct	91	**Minimum**	1916
Distinct (%)	1.8%	**Maximum**	2016
Missing	108	**Zeros**	0
Missing (%)	2.1%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2002.4705	**Memory size**	39.5 KiB

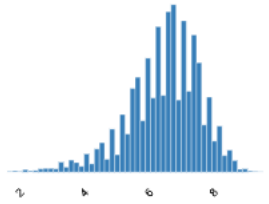


More details

- The average imdb rating for the listed movies was 6.44.

imdb_score
Real number (ℝ)

Distinct	78	**Minimum**	1.6
Distinct (%)	1.5%	**Maximum**	9.5
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6.4421376	**Memory size**	39.5 KiB



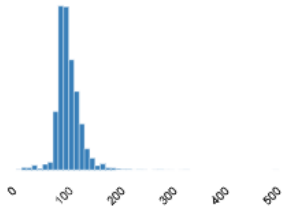
More details

- Movies included had an average duration of 107 minutes.

duration

Real number (ℝ)

Distinct	191	**Minimum**	7
Distinct (%)	3.8%	**Maximum**	511
Missing	15	**Zeros**	0
Missing (%)	0.3%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	107.20107	**Memory size**	39.5 KiB



Feature Engineering

Feature engineering helps to improve the predictive power of machine learning models. By creating new columns, we are able to extract more information from existing data and provide additional context for analysis.

- One hot encoding was applied on the categorical variables. One-hot encoding is a technique used to convert categorical variables into numerical variables that can be used in machine learning models. By creating these new numerical variables, we can provide the model with richer information. This will enable it to identify patterns and relationships that would be difficult to detect with raw categorical data.

- Standard scaling, also known as Z-score normalization, is a vital preprocessing step when working with numerical variables in machine learning and was applied to the imdb data. This technique involves transforming numerical features so that they have a mean of 0 and a standard deviation of 1. The importance of standard scaling lies in its ability to bring all numerical variables to a common scale, thus ensuring that no single feature dominates the learning process or model performance due to differences in their original ranges. When numerical features are measured in different units or have widely varying magnitudes, machine learning algorithms that rely on distance metrics (such as k-nearest neighbors or clustering) or gradient-based optimization (like linear regression or neural networks) can be adversely affected. Variables with larger scales might disproportionately influence the model's behavior, leading to biased parameter estimates and suboptimal convergence during training. Standard scaling addresses this issue by centering the data around its mean and then scaling it based on its standard deviation. This results in each feature having a similar scale, effectively placing them on equal footing. Consequently, machine learning algorithms can make more accurate comparisons and decisions, ensuring that no single feature dominates the others solely due to its original measurement range. Standard scaling also aids in speeding up convergence during optimization processes, contributing to more stable and reliable model performance.

Predictive Modeling

This report presents the application of three regression algorithms - K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting - to predict IMDb ratings of movies. The report outlines the chosen hyperparameters for each regressor and discusses the overall approach adopted for model development and evaluation.

Regressor Summaries and Hyperparameters:

1. K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm that predicts the target variable based on the majority class of its k nearest neighbors in the feature space. The hyperparameters considered for KNN are as follows:

   - n_neighbors: Number of neighbors to consider (3, 5, 7)
   - weights: Weight function used in prediction ('uniform', 'distance')
   - p: Distance metric used ('Manhattan' or 'Euclidean')

2. Random Forest: Random Forest is an ensemble technique that combines multiple decision trees to improve predictive performance. The hyperparameters explored for Random Forest are:

   - n_estimators: Number of trees in the forest (50, 100)
   - max_depth: Maximum depth of the trees (None, 10, 20)
   - min_samples_split: Minimum samples required to split an internal node (2, 5, 10)
   - min_samples_leaf: Minimum number of samples required to be at a leaf node (1, 2, 4)
   - max_features: Maximum number of features to consider for split ('auto', 'sqrt', 'log2')

3. Gradient Boosting: Gradient Boosting is an ensemble technique that sequentially builds weak learners to create a strong predictive model. The hyperparameters examined for Gradient Boosting are:
   - n_estimators: Number of boosting stages (50, 100)
   - learning_rate: Step size at each iteration (0.01, 0.1, 0.2)
   - max_depth: Maximum depth of the individual trees (3, 5, 7)
   - subsample: Fraction of samples used for fitting trees (0.8, 0.9, 1.0)
   - min_samples_split: Minimum samples required to split an internal node (2, 5, 10)

   Modeling Approach:

1. Data Preprocessing: The IMDb movie dataset was preprocessed, including handling missing values, encoding categorical variables, and splitting the dataset into training and testing sets (70-30 ratio).
2. Regressor Implementation: Three regression algorithms - KNN, Random Forest, and Gradient Boosting - were selected for modeling IMDb ratings.

3. Hyperparameter Tuning: Hyperparameters were fine-tuned using an exhaustive grid search approach. For each regressor, a grid of hyperparameter values was defined, and cross-validation was performed to identify the optimal combination.
4. Model Training and Evaluation: The models were trained on the training set using optimized hyperparameters. The models' performance was evaluated on the testing set using metrics - Root Mean Squared Error (RMSE), and R-squared.
5. Model Comparison: The performance of each regressor was compared based on evaluation metrics to determine which algorithm yielded the most accurate IMDb rating predictions.

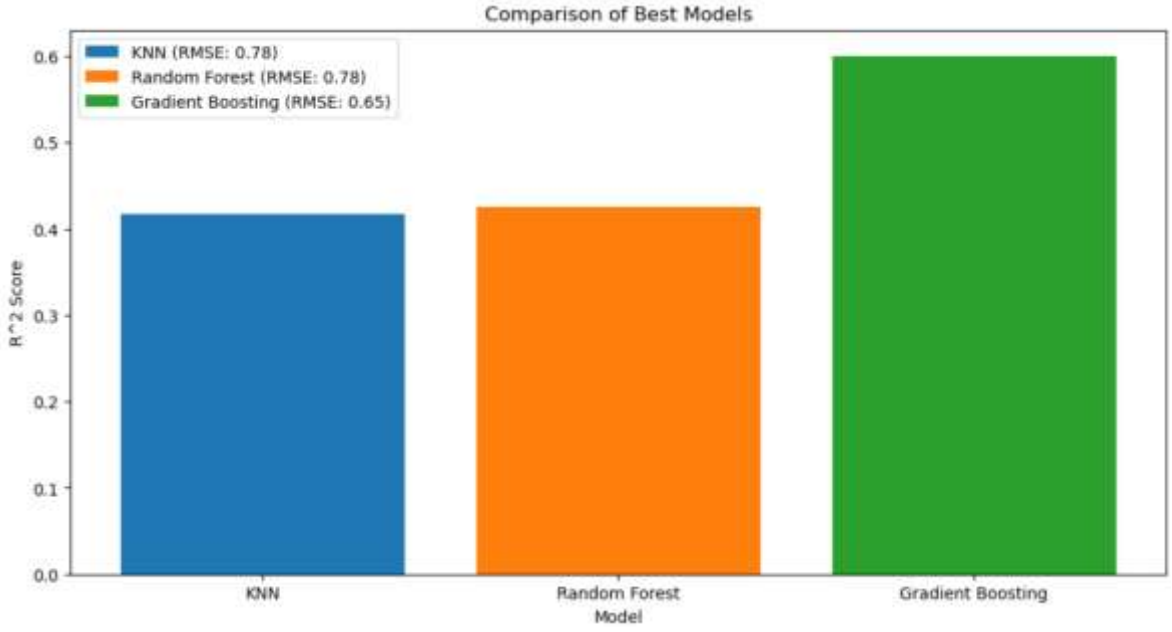
RECOMMENDED MODEL

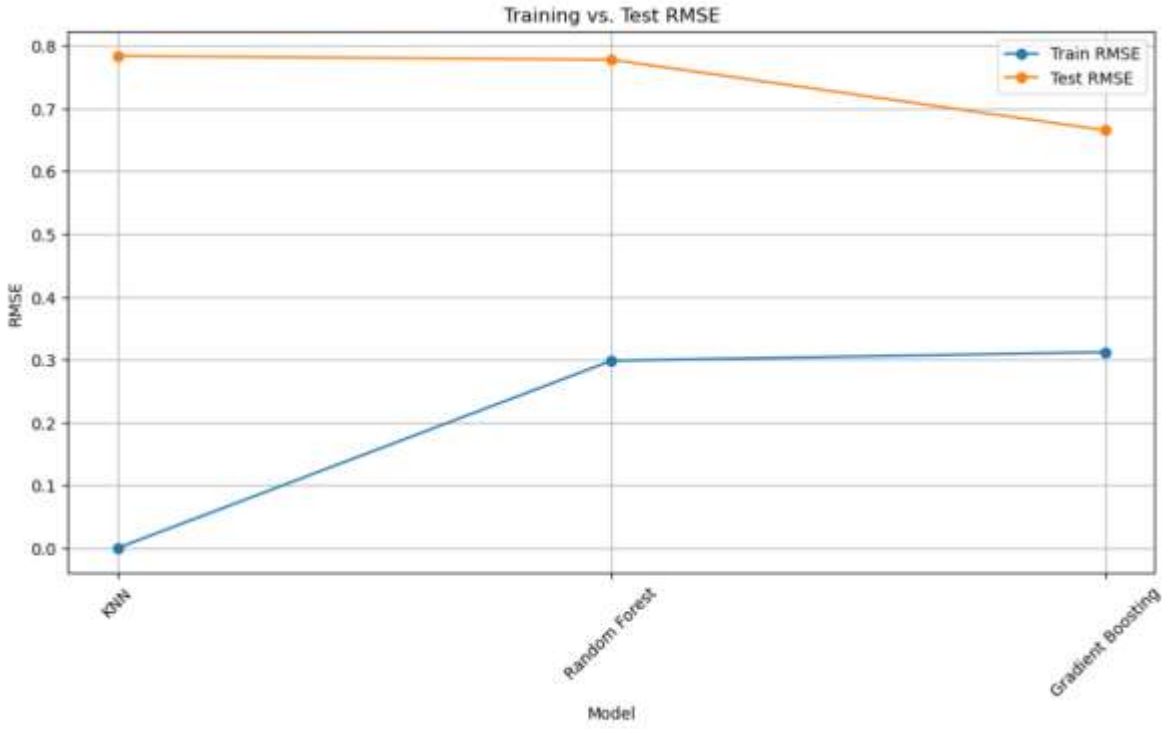
- **Model comparison**

Regressor	**R^2**	**RMSE**
KNN	0.41	0.78
Random Forest	0.42	0.79
Gradient Boosting	0.61	0.65

Model comparison across measurements

The models' performance was evaluated on the testing set, and the following summary provides insights into their predictive capabilities:





K-Nearest Neighbors (KNN):

Root Mean Squared Error (RMSE): 0.784
R-squared (R2) Score: 0.417

The KNN model demonstrates an ability to capture some of the variance in the IMDb ratings. However, the relatively high RMSE and relatively low R2 score indicate that the model's predictions have a moderate level of error and might not fully explain the variance in the target IMDb ratings. This could be attributed to the simple nature of the algorithm and potential limitations in capturing complex relationships within the dataset.

Random Forest:

Root Mean Squared Error (RMSE): 0.779
R-squared (R2) Score: 0.425

The Random Forest model exhibits slightly improved performance compared to KNN. The lower RMSE and marginally higher R2 score suggest that the ensemble nature of Random Forest allows it to capture more intricate relationships within the data. Nevertheless, there's room for further improvement, as indicated by the remaining prediction error and the potential for better explaining the variance in IMDb ratings.

Gradient Boosting:

Root Mean Squared Error (RMSE): 0.649
R-squared (R2) Score: 0.601

Among the three models, Gradient Boosting stands out as the most accurate predictor. With the lowest RMSE and the highest R2 score, the model effectively minimizes prediction errors and explains a substantial portion of the variance in IMDb ratings. The ensemble approach of sequentially boosting weak learners has enabled Gradient Boosting to capture intricate patterns and relationships within the data, resulting in superior predictive performance.

Conclusion

In summary, the evaluation results of the three regression models shed light on their respective capabilities in predicting IMDb movie ratings. While KNN and Random Forest exhibit moderate performance, the Gradient Boosting model stands out as the most accurate and effective predictor. Its ability to minimize RMSE and achieve a high R2 score demonstrates its proficiency in capturing complex relationships within the dataset, making it a strong candidate for accurate IMDb rating predictions.

Recommendations:

Recommendations for Enhancing IMDb Rating Prediction and Analysis are –

1. **Collect More Data and Address Class Imbalance:** To improve the predictive accuracy of IMDb movie ratings, it is recommended to gather a more extensive and diverse dataset. A larger dataset enables the models to capture a wider range of movie characteristics and trends, reducing the risk of overfitting and enhancing generalization. Additionally, addressing class imbalance within the dataset is crucial. Ensuring a representative distribution of IMDb rating classes can help models accurately predict ratings across the entire spectrum, from low to high ratings. This balanced representation contributes to more robust and reliable predictions.

2. **Leverage Social Media Data for Holistic Analysis:** The movie industry should consider integrating IMDb rating predictions with data from other social media platforms. In today's digital age, audiences share opinions and engage with movies across various online platforms, including social media networks, forums, and review websites. Integrating this social media data with IMDb ratings can provide a more holistic and comprehensive understanding of a movie's reception and impact. Sentiment analysis and topic modeling techniques can extract valuable insights from these diverse sources, helping filmmakers gain a deeper understanding of audience preferences and reactions.

3. **Incorporate Domain-Specific Features:** To enhance the accuracy of IMDb rating predictions, it is recommended to include domain-specific features that have a significant impact on movie ratings. These features could include directorial experience, cast popularity, production budget, genre trends, release timing, and more. By incorporating these factors, the models can better capture the nuances that contribute to a movie's reception, resulting in more accurate predictions.

4. **Regular Model Maintenance and Reevaluation:** As the movie industry is dynamic and ever evolving, it's essential to continuously update and refine the predictive models. Regularly incorporating new data, adjusting hyperparameters, and exploring advanced modeling techniques can ensure that the models remain relevant and effective. This adaptability allows the movie industry to stay on top of emerging trends and changing audience preferences.

5. **Collaborative Industry Insights:** The movie industry should encourage collaboration between data scientists, filmmakers, producers, and other industry stakeholders. This interdisciplinary approach ensures that the predictive models are aligned with the industry's practical needs and objectives. Industry experts can provide valuable insights into relevant features, potential biases, and critical decision-making factors that should be considered during model development and analysis.

*** End of Report ***