Predicting IMDb Scores: Leveraging Pre-Release Movie Attributes

Jaswanth Kranthi Boppana   Vishnuvardhan Reddy Kollu   Vineela Kunisetti

Abstract

This project presents a predictive analytics framework for anticipating IMDb scores of upcoming movies using pre-release data from the IMDb 5000 dataset. By preprocessing and refining the dataset to include only attributes available before a movie's release, such as director reputation, genre, and social media engagement etc, we create a robust foundation for machine learning analysis. Regression models, including LightGBM, XGBoost, and Gradient Boosting, predict continuous IMDb scores, while classification models categorize movies into predefined rating brackets. Feature importance analysis identifies key drivers of ratings, enabling actionable insights into audience preferences and market dynamics. The results are further enhanced using GPT to generate human-readable interpretations, bridging technical outputs with practical recommendations for filmmakers and producers. This integrated approach provides a comprehensive and accessible system to guide creative and strategic decisions, aligning production efforts with audience expectations to optimize movie success.

Keywords: IMDb Score, Predictive modelling, Exploratory Data Analytics, Data Preprocessing, Feature Engineering, Regression, classification and data-driven insights.

I. INTRODUCTION

The global film industry, valued at approximately $136 billion in 2022 [1], is increasingly driven by data and analytics to optimize production and marketing strategies. Platforms like IMDb, attracting over 200 million monthly visitors [2], play a pivotal role in shaping audience perceptions and success indicators for films. Predictive modelling offers filmmakers and producers an opportunity to make informed decisions before a movie's release, leveraging pre-release data to anticipate audience reception. This study focuses on utilizing attributes such as director reputation, genre, and social media engagement to predict IMDb scores, addressing significant gaps in existing methodologies.

Erik Bruin's Kaggle project demonstrated the use of IMDb ratings, genres, and user inputs to develop recommendation systems [3]. While effective in improving user engagement, the approach relies heavily on post-release user interactions, rendering it unsuitable for pre-release prediction. Similarly, Joseph et al. (2020) classified IMDb scores into success categories but did not exclusively utilize pre-release attributes and omitted critical metrics such as social media presence and actor influence [4]. Addressing these limitations, our study focuses on pre-release attributes to generate actionable insights.

Additional contributions to this field include Jose and Harikumar's (2022) work, which employed RoBERTa embeddings and neural networks to predict IMDb ratings based on storyline embeddings and genre [5]. While capturing narrative-driven insights, the study lacked integration of broader pre-release factors such as social media engagement and director reputation. Similarly, an arXiv study (2024) explored movie revenue prediction but centred on financial metrics rather than audience perception [6]. A study by T. Kumar et al. (2020) on IEEE utilized neural networks for IMDb score prediction, combining pre-release attributes such as genre, release date, and runtime [7]. However, it did not incorporate metrics like social media buzz or historical performance of directors and actors, which are key drivers of success.

Our approach bridges these gaps by employing pre-release attributes—including aggregated metrics of director reputation, production details, and social media engagement—in a comprehensive predictive model. Using regression techniques like LightGBM and XGBoost, we predict IMDb scores as a continuous variable and categorize movies into rating brackets through classification models. Feature importance analysis enables us to identify critical factors influencing outcomes, and we utilize GPT models to translate these findings into human-readable insights. This holistic framework not only improves predictive accuracy but also provides filmmakers and producers with actionable strategies to align their creative and marketing decisions with audience expectations.

II. Dataset Details

The dataset for this project comprises 5043 records spanning over 100 years and covering 66 countries, providing a robust foundation for predicting IMDb scores for movies. It includes 28 variables, with imdb_score as the target and 27 predictor variables encompassing attributes like director names, movie genres, budgets, gross earnings, and audience interactions. Performance-related metrics such as the number of reviews, Facebook likes for cast and crew, and plot keywords offer valuable insights into pre-release appeal and potential success.

This dataset's mix of numerical and categorical variables enables a comprehensive analysis of factors influencing a movie's reception. High-cardinality variables, such as director and actor names, were transformed into aggregated historical metrics to ensure meaningful analysis. With its depth and diversity, the dataset supports predictive modelling and provides actionable insights into the drivers of movie success.

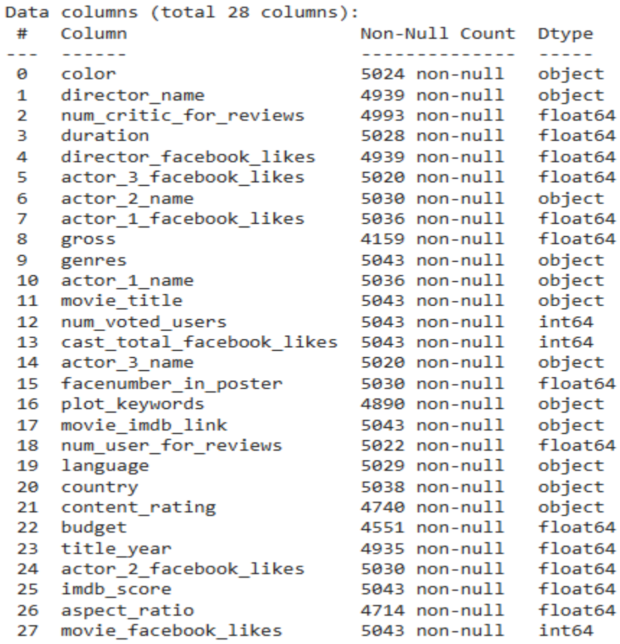
Key Features in the Dataset:

- **Director Significance:** Directors' historical performance, reflected through their average IMDb scores, emerges as a strong indicator of a movie's potential success. Their influence on creative and production aspects makes this a critical feature in predicting audience and critical reception.

- **Production Attributes:** Key production-related factors, such as budget, highlight the importance of financial investment in ensuring high production quality, which directly impacts a movie's IMDb score. This underscores the need for strategic resource allocation during the planning stages.
- **Audience Interaction Metrics:** Variables like the number of user and critic reviews provide valuable insights into audience engagement and critical validation. These metrics indicate the degree of interest and approval a movie generates before and after its release.
- **Social Media Presence:** Facebook likes for directors, cast members, and movies act as a proxy for digital engagement, demonstrating the growing role of social media in creating pre-release buzz. This highlights the importance of a strong digital marketing strategy.
- **Aggregated Historical Data:** Metrics derived from historical performance, such as average IMDb scores for directors, offer a reliable benchmark for predicting the potential reception of new projects. This aggregated data streamlines high-cardinality variables and ensures a robust predictive framework.

III. Data Preprocessing

Data preprocessing is a critical step in transforming raw data into actionable insights, particularly for predictive analytics. This phase involves multiple tasks, including handling missing values, encoding categorical variables, scaling numerical features, and performing exploratory data analysis (EDA). Each component of this process is essential for refining the dataset, ensuring compatibility with machine learning algorithms, and uncovering meaningful patterns to enhance predictive accuracy and decision-making.



```
Data columns (total 28 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   color                      5024 non-null   object
 1   director_name              4939 non-null   object
 2   num_critic_for_reviews     4993 non-null   float64
 3   duration                   5028 non-null   float64
 4   director_facebook_likes    4939 non-null   float64
 5   actor_3_facebook_likes     5020 non-null   float64
 6   actor_2_name               5030 non-null   object
 7   actor_1_facebook_likes     5036 non-null   float64
 8   gross                      4159 non-null   float64
 9   genres                     5043 non-null   object
 10  actor_1_name               5036 non-null   object
 11  movie_title                5043 non-null   object
 12  num_voted_users            5043 non-null   int64
 13  cast_total_facebook_likes  5043 non-null   int64
 14  actor_3_name               5020 non-null   object
 15  facenumber_in_poster       5030 non-null   float64
 16  plot_keywords              4890 non-null   object
 17  movie_imdb_link            5043 non-null   object
 18  num_user_for_reviews       5022 non-null   float64
 19  language                   5029 non-null   object
 20  country                    5038 non-null   object
 21  content_rating             4740 non-null   object
 22  budget                     4551 non-null   float64
 23  title_year                 4935 non-null   float64
 24  actor_2_facebook_likes     5030 non-null   float64
 25  imdb_score                 5043 non-null   float64
 26  aspect_ratio               4714 non-null   float64
 27  movie_facebook_likes       5043 non-null   int64
```

Figure 1: Dataset Attributes Before Preprocessing

Handling missing values:

Missing data was handled carefully to maintain the integrity of the dataset while retaining important features. Columns with excessive missing values, such as gross and budget, were filtered to remove rows with null entries, as imputation for these variables would not yield accurate results. For other columns, missing values were replaced with meaningful substitutes; for example, facenumber_in_poster was filled with its mean, and crew-related Facebook likes were filled with the average values based on grouped data. Additionally, missing content_rating values were replaced with the most frequent category to maintain consistency.

Data Encoding :

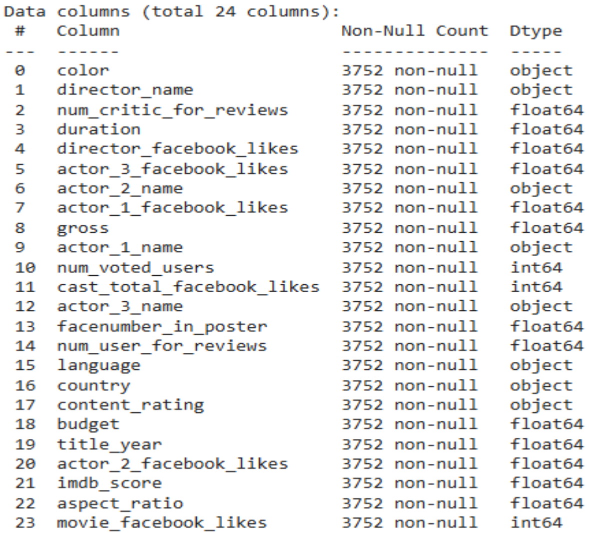
Categorical variables were encoded to transform them into numerical formats suitable for machine learning. For low-cardinality variables like color, language, country, and content_rating, one-hot encoding was applied to create binary indicator columns. High-cardinality variables, such as director_name and actor_1_name, were target-encoded by calculating the mean IMDb score for each unique value, effectively condensing large categories into meaningful numeric representations. These steps ensured the dataset was both interpretable and compact.

Outlier Detection and Removal:

Extreme outliers in key numerical variables, such as budget and gross, were identified and removed using z-score analysis. Rows with z-scores exceeding a threshold of ±3 were excluded, ensuring that outliers did not skew the results or mislead the analysis.

Data Simplification:

Less relevant columns, such as movie_imdb_link, plot_keywords, and movie_title, were removed to reduce noise and focus on impactful features. Furthermore, the country variable was simplified into three categories: USA, UK, and Others, to make the analysis more manageable while retaining key distinctions. By following these preprocessing steps, the dataset was transformed into aclean,standardized, and feature-rich structure, enabling the application of robust machine learning models for predicting IMDb scores.



```
Data columns (total 24 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   color                      3752 non-null   object
 1   director_name              3752 non-null   object
 2   num_critic_for_reviews     3752 non-null   float64
 3   duration                   3752 non-null   float64
 4   director_facebook_likes    3752 non-null   float64
 5   actor_3_facebook_likes     3752 non-null   float64
 6   actor_2_name               3752 non-null   object
 7   actor_1_facebook_likes     3752 non-null   float64
 8   gross                      3752 non-null   float64
 9   actor_1_name               3752 non-null   object
 10  num_voted_users            3752 non-null   int64
 11  cast_total_facebook_likes  3752 non-null   int64
 12  actor_3_name               3752 non-null   object
 13  facenumber_in_poster       3752 non-null   float64
 14  num_user_for_reviews       3752 non-null   float64
 15  language                   3752 non-null   object
 16  country                    3752 non-null   object
 17  content_rating             3752 non-null   object
 18  budget                     3752 non-null   float64
 19  title_year                 3752 non-null   float64
 20  actor_2_facebook_likes     3752 non-null   float64
 21  imdb_score                 3752 non-null   float64
 22  aspect_ratio               3752 non-null   float64
 23  movie_facebook_likes       3752 non-null   int64
```

Figure 2: Dataset Attributes After Preprocessing

IV. Feature Engineering

Feature engineering plays a pivotal role in this project due to the nature of the available dataset. Since our focus is on utilizing metrics and features accessible **before a movie's release**, we encountered certain post-release features within the dataset. However, these post-release features showed significant correlations with pre-release attributes, allowing us to derive new, meaningful features that can be utilized in IMDb score prediction. Below, we detail the creation of two key engineered features:

Expected Engagement

To evaluate audience and reviewer engagement with a movie's director, we engineered the feature *expected_engagement*. This metric was developed by combining the attributes `num_critic_for_reviews`, `num_voted_users`, and `num_user_for_reviews`. These features were scaled to a uniform range of 1 to 10 using the MinMaxScaler and aggregated by `director_name`. The resulting variable quantifies a director's ability to generate engagement from audiences and critics, offering a pre-release proxy for audience interest and anticipation.

Director's Average IMDb (Directors Avg_IMDb)

Recognizing the critical influence of directors on a movie's success, we introduced a new feature called *Directors Avg_IMDb*. This metric represents the average IMDb ratings of a director's previous movies. By aggregating and calculating historical IMDb ratings for each director, we provided a robust, data-driven evaluation of their track record, which serves as a strong predictor of future movie performance.

Standardization of Numerical Features

To enhance model performance and stability, numerical variables such as `budget`, `gross`, and `Facebook likes` were standardized using a StandardScaler. This ensured that all numerical features operated on a similar scale, reducing potential bias during model training and improving the consistency of predictions.

V. EDA

After **feature engineering**, Exploratory Data Analysis (EDA) was performed to uncover patterns and relationships within the dataset. Visualizations, including scatter plots, correlation matrices, and time series trends, revealed insights such as the positive link between user votes and IMDb scores and evolving trends in budgets and gross earnings over time. These findings guided the selection of key features, ensuring the models were built on the most relevant and impactful data.

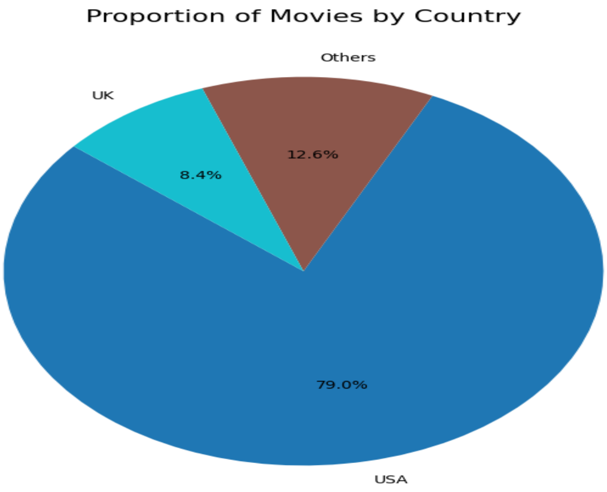


Fig 3: Distribution of Movies by Country

The pie chart in Fig 3 shows that the majority of movies in the dataset (79%) are produced in the USA, followed by the UK at 8.4%, and 12.6% from other countries. This highlights a strong dominance of the US film industry, which may influence trends in IMDb scores and global audience engagement.

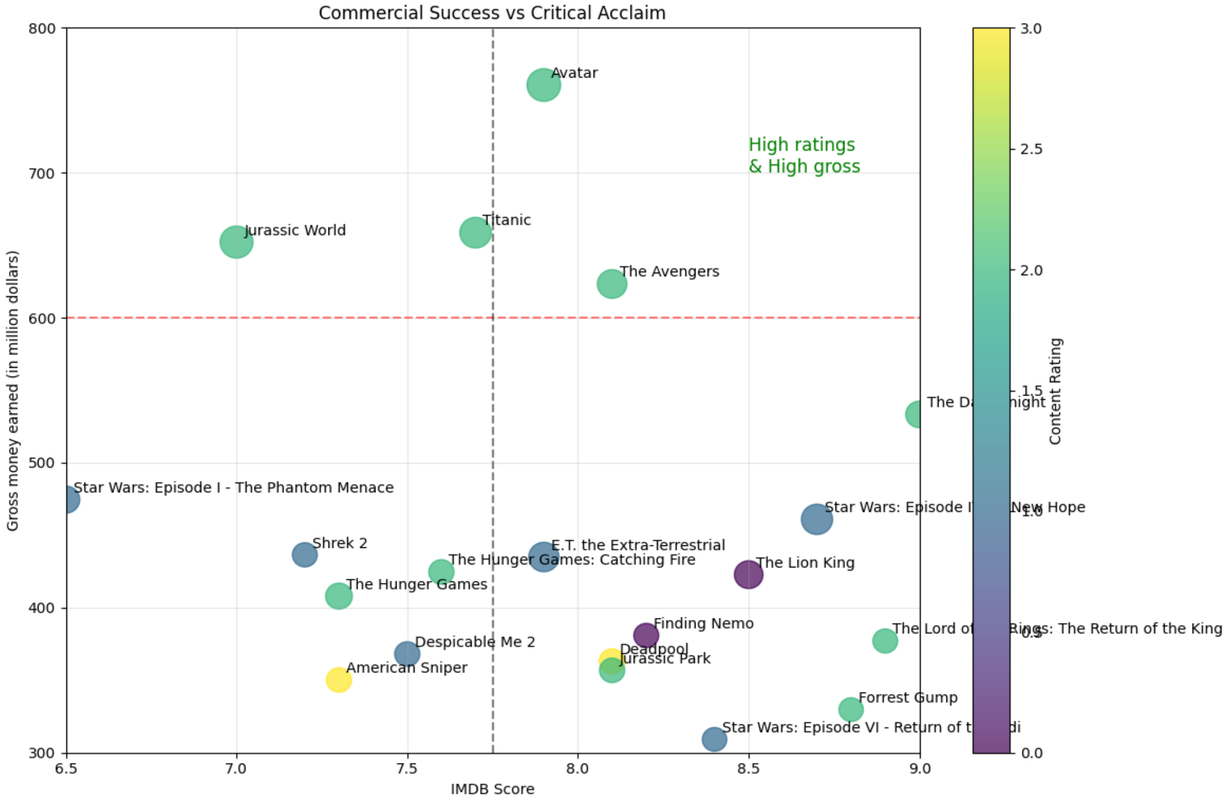


Fig 4: Commercial Success vs Critical Acclaim

The bubble chart in Fig 4 illustrates the relationship between IMDb scores (critical acclaim) and gross earnings (commercial success) for popular movies. Movies like *Avatar* and *Titanic* achieved both high IMDb ratings and substantial box office earnings, placing them in the "High Ratings & High Gross" quadrant. However, movies with lower IMDb scores, such as *Shrek 2* and *The Phantom Menace*, still demonstrated significant commercial success. This analysis highlights that while high ratings and financial success often align, certain films can achieve strong box office performance despite moderate critical acclaim, indicating other factors like audience appeal and marketing strategies.

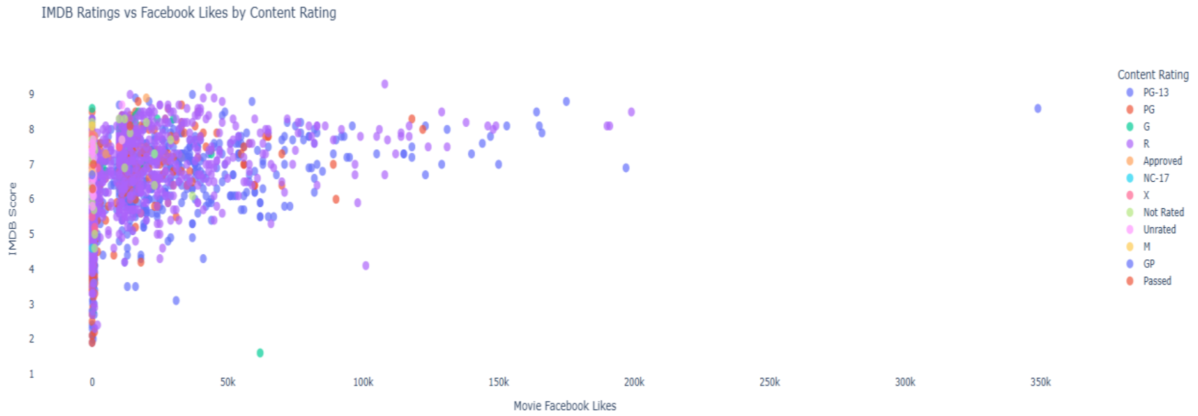
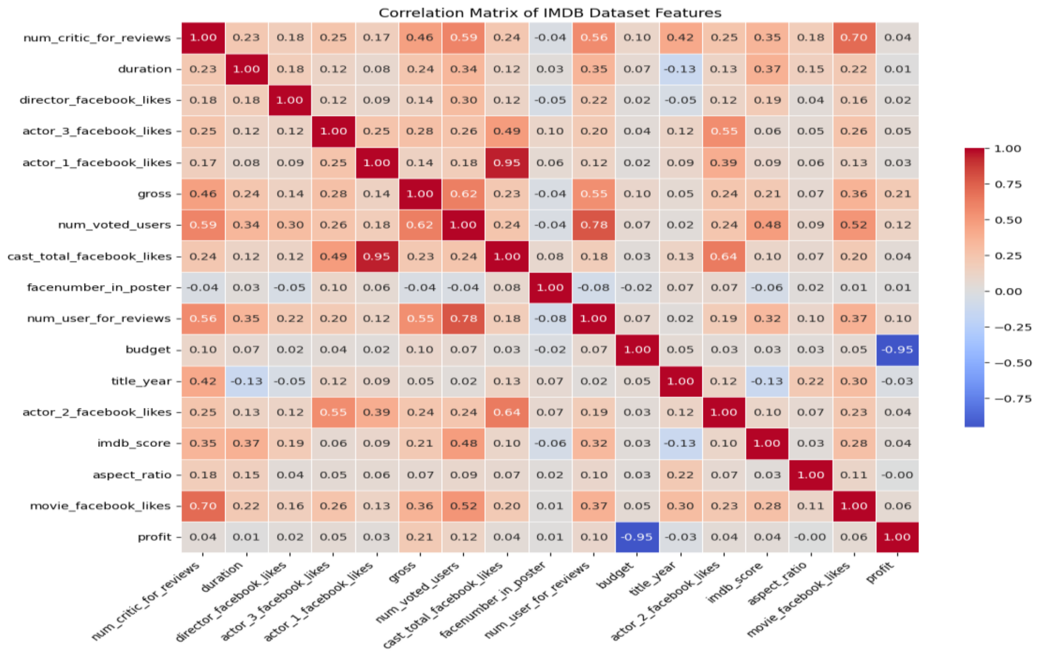


Fig 5: IMDb Ratings vs Facebook Likes by Content Rating

The scatter plot in Fig 5 explores the relationship between IMDb scores and movie Facebook likes, categorized by content ratings. It reveals that higher IMDb scores tend to align with moderate to high Facebook engagement, although there is significant variability. Movies with ratings like PG-13 and R dominate the higher engagement range, suggesting broader audience appeal and popularity. This highlights the influence of social media presence on audience perception and pre-release hype.



	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes	gross	num_voted_users	cast_total_facebook_likes	facenumber_in_poster	num_user_for_reviews	budget	title_year	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes	profit
num_critic_for_reviews	1.00	0.23	0.18	0.25	0.17	0.46	0.59	0.24	-0.04	0.56	0.10	0.42	0.25	0.35	0.18	0.70	0.04
duration	0.23	1.00	0.18	0.12	0.08	0.24	0.34	0.12	0.03	0.35	0.07	-0.13	0.13	0.37	0.15	0.22	0.01
director_facebook_likes	0.18	0.18	1.00	0.12	0.09	0.14	0.30	0.12	-0.05	0.22	0.02	-0.05	0.12	0.19	0.04	0.16	0.02
actor_3_facebook_likes	0.25	0.12	0.12	1.00	0.25	0.28	0.26	0.49	0.10	0.20	0.04	0.12	0.55	0.06	0.05	0.26	0.05
actor_1_facebook_likes	0.17	0.08	0.09	0.25	1.00	0.14	0.18	0.95	0.06	0.12	0.02	0.09	0.39	0.09	0.06	0.13	0.03
gross	0.46	0.24	0.14	0.28	0.14	1.00	0.62	0.23	-0.04	0.55	0.10	0.05	0.24	0.21	0.07	0.36	0.21
num_voted_users	0.59	0.34	0.30	0.26	0.18	0.62	1.00	0.24	-0.04	0.78	0.07	0.02	0.24	0.48	0.09	0.52	0.12
cast_total_facebook_likes	0.24	0.12	0.12	0.49	0.95	0.23	0.24	1.00	0.08	0.18	0.03	0.13	0.64	0.10	0.07	0.20	0.04
facenumber_in_poster	-0.04	0.03	-0.05	0.10	0.06	-0.04	-0.04	0.08	1.00	-0.08	-0.02	0.07	0.07	-0.06	0.02	0.01	0.01
num_user_for_reviews	0.56	0.35	0.22	0.20	0.12	0.55	0.78	0.18	-0.08	1.00	0.07	0.02	0.19	0.32	0.10	0.37	0.10
budget	0.10	0.07	0.02	0.04	0.02	0.10	0.07	0.03	-0.02	0.07	1.00	0.05	0.03	0.03	0.03	0.05	-0.95
title_year	0.42	-0.13	-0.05	0.12	0.09	0.05	0.02	0.13	0.07	0.02	0.05	1.00	0.12	-0.13	0.22	0.30	-0.03
actor_2_facebook_likes	0.25	0.13	0.12	0.55	0.39	0.24	0.24	0.64	0.07	0.19	0.03	0.12	1.00	0.10	0.07	0.23	0.04
imdb_score	0.35	0.37	0.19	0.06	0.09	0.21	0.48	0.10	-0.06	0.32	0.03	-0.13	0.10	1.00	0.03	0.28	0.04
aspect_ratio	0.18	0.15	0.04	0.05	0.06	0.07	0.09	0.07	0.02	0.10	0.03	0.22	0.07	0.03	1.00	0.11	-0.00
movie_facebook_likes	0.70	0.22	0.16	0.26	0.13	0.36	0.52	0.20	0.01	0.37	0.05	0.30	0.23	0.28	0.11	1.00	0.06
profit	0.04	0.01	0.02	0.05	0.03	0.21	0.12	0.04	0.01	0.10	-0.95	-0.03	0.04	0.04	-0.00	0.06	1.00

Fig 6: Correlation Matrix of IMDb Dataset Features

The correlation matrix in Fig 6 provides a detailed view of relationships between numerical features in the IMDb dataset. Notable positive correlations include num_critic_for_reviews and gross (0.46), as well as num_voted_users and imdb_score (0.48), indicating that higher audience and critic engagement align with better movie performance. Budget and gross also show moderate positive correlation (0.62), reflecting the role of financial investment in achieving commercial success.

Interestingly, the budget and profit correlation is nearly negligible, indicating that higher budgets do not always guarantee proportional returns. Additionally, features like cast_total_facebook_likes and actor_1_facebook_likes are strongly correlated (0.95), suggesting redundancy between these variables. Such observations from the correlation matrix guided feature selection and preprocessing to improve model performance.

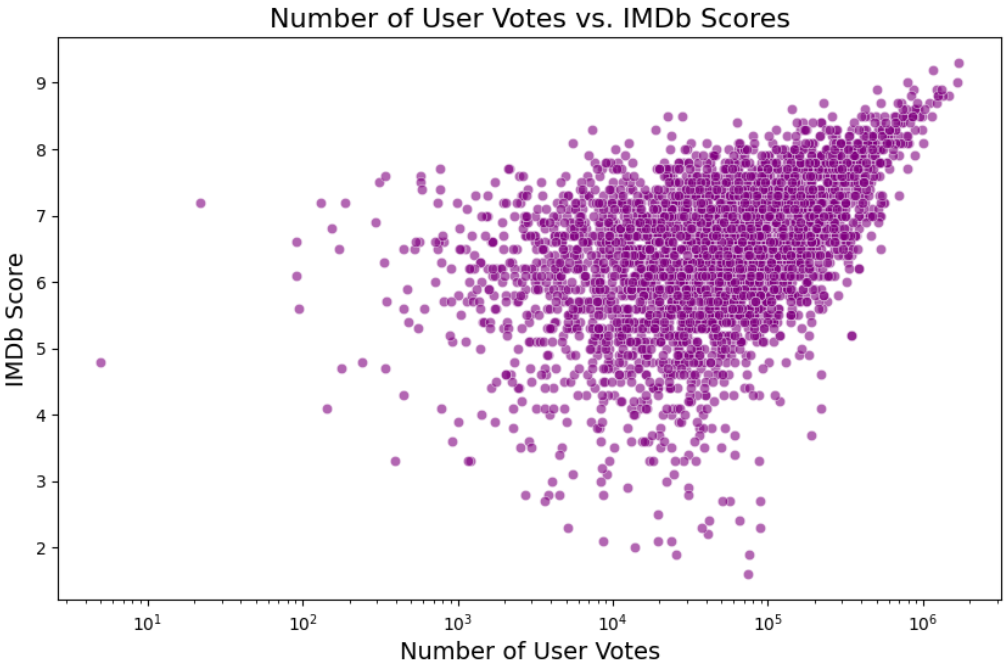


Fig 7: Relationship Between User Votes and IMDb Scores

The scatter plot in Fig 7 shows a positive relationship between the number of user votes and IMDb scores. Movies with higher user votes generally tend to have higher IMDb scores, suggesting that popular movies often receive favourable ratings, reinforcing the link between audience engagement and critical reception.

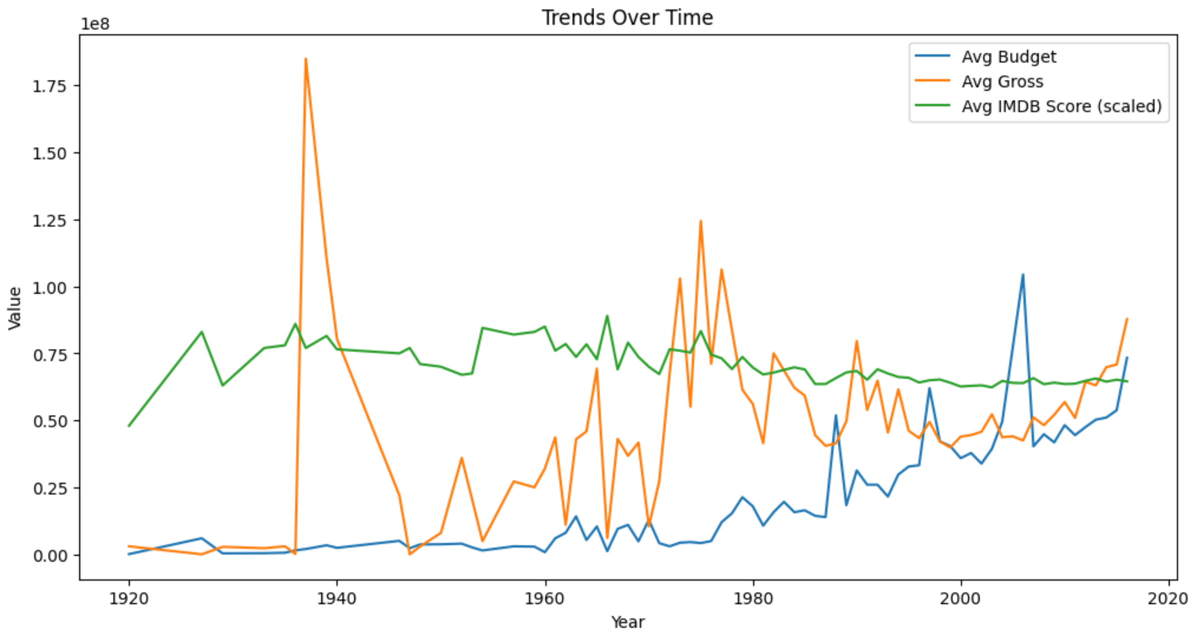


Fig 8: Trends in Average Budget, Gross, and IMDb Scores Over Time

The line chart in Fig 8 shows trends in average budget, gross earnings, and IMDb scores over time. While budgets and gross earnings have increased significantly in recent decades, IMDb scores have remained relatively stable. This indicates that higher production costs do not necessarily translate to better critical reception.

VI. Methodology

After cleaning the data and creating the best possible features through feature engineering, we employed a diverse range of machine learning models, strategically selected to leverage their unique strengths in predictive accuracy and interpretability. These techniques were categorized into regression and classification approaches, each tailored to different aspects of the analysis. Regression models, including LightGBM, XGBoost, Random Forest, Linear Regression, and Gradient Boosting, were used to predict continuous IMDb scores with precision. Meanwhile, classification models, such as LightGBM, Random Forest, and Gradient Boosting, were deployed to categorize movies into predefined rating brackets, offering a complementary perspective on movie success. This section explores the implementation and impact of these techniques, emphasizing their role in delivering robust and actionable insights. The represents the flow of the methodology.

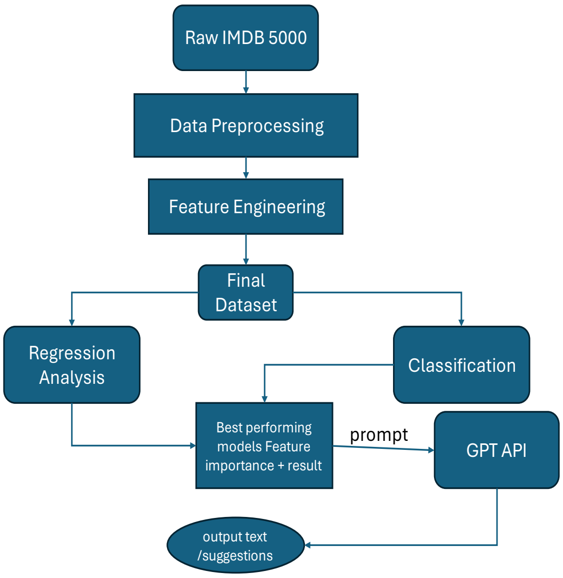


Fig 9: Methodology Flow chat

To begin, we start with the regression analysis. The dataset is split into training and testing sets with a split ratio of 80% and 20%, respectively. Various regression models are trained using the training data, and the loss metrics are calculated. The following content describes the techniques that were employed.

Regression Techniques

1. **LightGBM**

   LightGBM (Light Gradient Boosting Machine) builds trees leaf-wise, allowing it to efficiently handle large datasets and capture non-linear relationships. Its ability to manage categorical features and reduce computation time makes it ideal for datasets with high-dimensional attributes, like movie metadata. In this experiment, the model achieved an R^2 value of **0.7557**, with parameters including 100 estimators, a maximum depth of 5, and a learning rate of 0.1. These settings balance model complexity and accuracy while minimizing overfitting.

2. **XGBoost**

   XGBoost (Extreme Gradient Boosting) is known for its scalability and ability to manage missing data, making it well-suited for predicting IMDb scores from potentially incomplete or sparse pre-release attributes. The model achieved an R^2 value of **0.7531**. It was configured with 100 estimators, a maximum depth of 5, and a learning rate of 0.1. XGBoost's regularization features, like L1 and L2 penalties, also help prevent overfitting, which is crucial for accurate predictions.

3. **Gradient Boosting**

   Gradient Boosting builds trees sequentially, focusing on correcting errors made by previous trees. Its robustness in capturing intricate patterns makes it a reliable choice for predicting IMDb scores based on complex interactions among movie attributes. With an R2R^2R2 value of **0.7516**, this model used 100 estimators, a maximum depth of 5, and a learning rate of 0.1. While slightly slower than newer implementations like LightGBM, it remains an effective and interpretable choice.

4. **Random Forest**

   Random Forest, an ensemble method that averages the predictions of multiple decision trees, is highly robust to overfitting and works well with datasets containing a mix of numerical and categorical variables. Its R2R^2R2 value of **0.7148** suggests it captured general patterns in the data, though it struggled with more nuanced relationships. The model used 100 estimators and a maximum depth of 10, leveraging its capacity to handle non-linearities in the data.

5. **Linear Regression**

   Linear Regression serves as a baseline, assuming a linear relationship between movie attributes and IMDb scores. Its simplicity provides a starting point for understanding the data's linear trends. The model achieved an R2R^2R2 value of **0.6807**, highlighting its inability to fully capture the non-linearities and complex interactions inherent in movie metadata. Despite this, it remains valuable for identifying key linear contributors to IMDb scores.

The detailed analysis above highlights the unique strengths of each regression model and their suitability for predicting IMDb scores based on pre-release movie attributes. The clear performance comparison of each model, including their evaluation metrics such as R^2, MAE, and RMSE, is thoroughly presented in the results section, providing a comprehensive understanding of their effectiveness in this task.

Classification Method:

We have also implemented a classification method mainly because in the generic scenario a specific imdb score doesn't impact much but the range in which the score is in impacts the decision of the film. So we felt its important to try classification and segregate the movies into classes. In this classification method, we aim to predict movie IMDb scores by converting the continuous target variable `imdb_score` into four distinct classes:

Poor (0-4)
Average (4-6)
Good (6-8)
Excellent (8-10)

This classification approach allows for a more interpretable model output by categorizing movies into qualitative bins. However, upon initial analysis, a significant **class imbalance** was observed:

- **Good (6 - 8):** 2014 samples
- **Average (4 - 6):** 924 samples
- **Excellent (8 - 10):** 98 samples
- **Poor (0 - 4):** 80 samples

The majority of movies fell into the "Good (6 - 8)" range, comprising more than 64% of the dataset, while the "Poor (0 - 4)" and "Excellent (8 - 10)" categories were severely underrepresented, accounting for less than 3% each. To mitigate this imbalance, **Stratified Sampling** was employed to ensure that the distribution of classes remained consistent across training and testing splits.

The dataset consisted of categorical features like `language` and `country` which are neutralized by applying one-hot encoding, numerical attributes like `duration`, `gross`, and `budget` also exist These features represent a mix of pre-release movie attributes, which are critical indicators for predicting audience sentiment reflected through IMDb scores.

Models and Their Suitability:

To address the multi-class classification problem effectively, we explored three powerful machine learning models, chosen for their ability to handle structured data, non-linearity, and class imbalance:

1. **LightGBM Classifier**

   LightGBM is a gradient-boosting framework optimized for speed and performance, particularly on high-dimensional data. Its leaf-wise tree growth minimizes loss efficiently, and it supports handling class imbalance through parameters like `class_weight` or `is_unbalance`. Given the nature of our dataset, which includes numerical and encoded categorical variables, LightGBM is a natural choice due to its ability to handle sparse features and capture intricate patterns. LightGBM obtained an accuracy of 84.5%

2. **Random Forest Classifier**

   Random Forest is a robust ensemble method that builds multiple decision trees using bootstrap sampling and averages their predictions. Its ability to manage complex, non-linear relationships and reduce overfitting makes it well-suited for datasets with many features. Additionally, Random Forest provides feature importance, offering interpretability by identifying the most influential movie attributes for classification. On this dataset Random Forest Classifier obtained 81.7% accuracy.

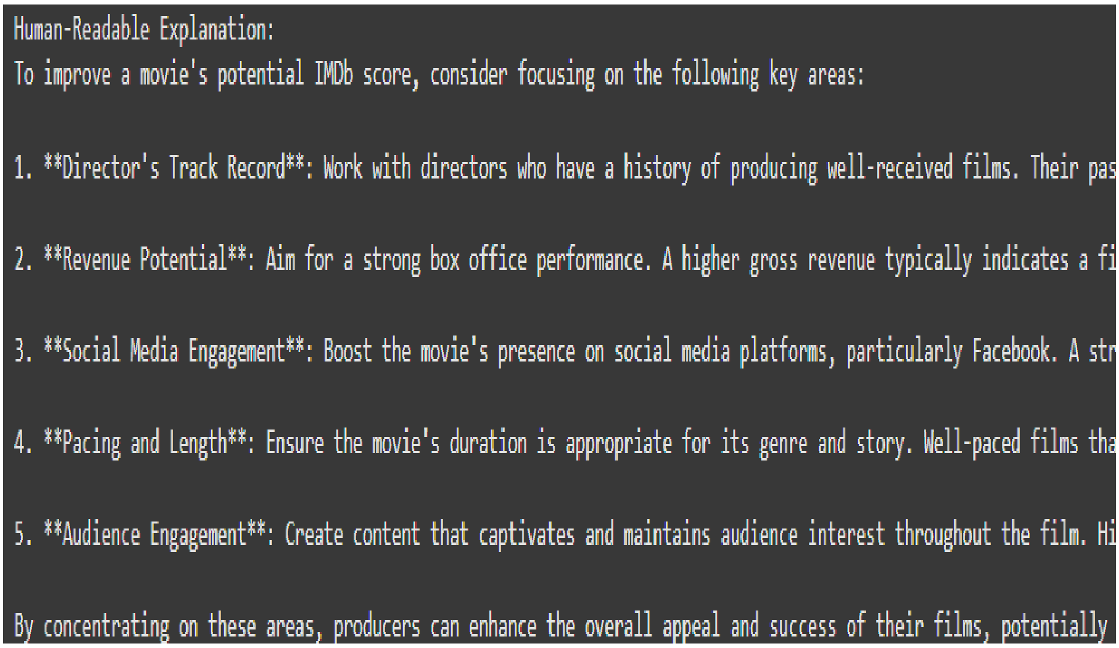
3. **Gradient Boosting Classifier**

   Gradient Boosting sequentially builds trees, optimizing for errors made by prior iterations. It is particularly effective at capturing subtle patterns in structured data and handling imbalanced classes when combined with appropriate resampling techniques or class weighting. Its flexibility allows for fine-tuning of hyperparameters such as learning rate and tree depth to optimize performance. Gradient Boosting gave accuracy of 83.9%.

   The comparative results of all classification models, including their performance metrics, are comprehensively presented and discussed in the results section, providing a clear understanding of their effectiveness in predicting IMDb score classes.

Integrating GPT to obtain Human readable suggestions:

In addition to performing the regression and classification and predicting the imdb score or the class of the imdb score we are intending to provide user readable suggestion to make to get a better imdb score. To do this we are extracting the feature importance of the best performing model and prompting the open AI GPT API with the feature importance and the result to obtain user specific suggestions.



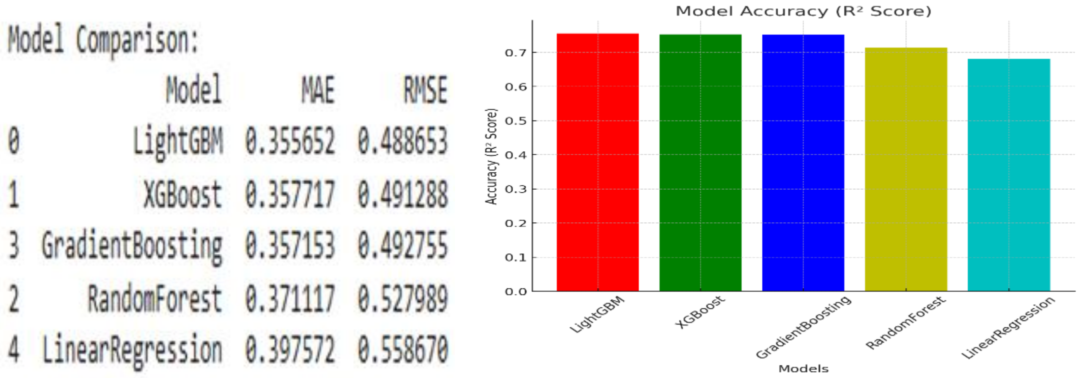
```
Human-Readable Explanation:
To improve a movie's potential IMDb score, consider focusing on the following key areas:

1. **Director's Track Record**: Work with directors who have a history of producing well-received films. Their pas

2. **Revenue Potential**: Aim for a strong box office performance. A higher gross revenue typically indicates a fi

3. **Social Media Engagement**: Boost the movie's presence on social media platforms, particularly Facebook. A str

4. **Pacing and Length**: Ensure the movie's duration is appropriate for its genre and story. Well-paced films tha

5. **Audience Engagement**: Create content that captivates and maintains audience interest throughout the film. Hi

By concentrating on these areas, producers can enhance the overall appeal and success of their films, potentially
```

Fig 10: Sample suggestion response from API

The text in Fig 10 shows the example output from the GPT after triggering the automation flow of prompting it with the feature importance metrics and the resulted output.

VII. Results

We would like to compare the performance on each model implemented in the regression analysis and the classification analysis. Later define the best performing model for this scenario.



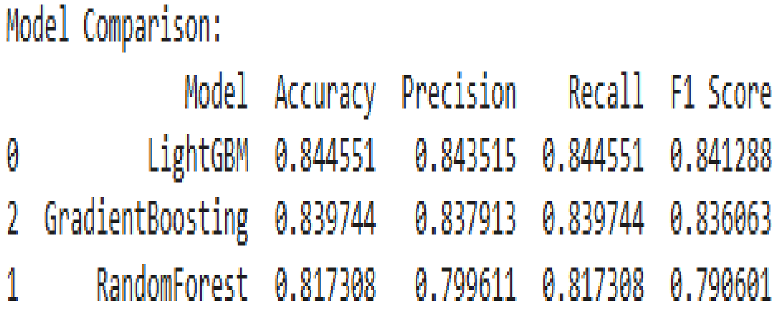
```
Model Comparison:
              Model       MAE      RMSE
0          LightGBM  0.355652  0.488653
1           XGBoost  0.357717  0.491288
3  GradientBoosting  0.357153  0.492755
2      RandomForest  0.371117  0.527989
4  LinearRegression  0.397572  0.558670
```

Fig 11: Regression Analysis Metrics

The above image shows the performance metrics of various methods are tested for predicting exact Imdb score. Form the image we can understand that the following

The LightGBM comes up as the top performer with the lowest MAE (0.355) and RMSE (0.488), and the highest R² (0.755), indicating it accurately captures the variability in IMDb scores. Whereas the XGBoost when Compared to LightGBM, with slightly higher RMSE and lower R², making it another strong candidate. Gradient Boosting Achieved results similar to XGBoost, with competitive performance in MAE and RMSE. whereas Random Forest Performed moderately well but lagged gradient boosting models due to higher RMSE and lower R². Linear Regression Struggled with the dataset's non-linear relationships, resulting in higher MAE and RMSE.

Now let’s see how the classification part of the models performed the dataset. Like regression, we have implemented various models and compared performance metrics.



```
Model Comparison:
              Model  Accuracy  Precision    Recall  F1 Score
0          LightGBM  0.844551   0.843515  0.844551  0.841288
2  GradientBoosting  0.839744   0.837913  0.839744  0.836063
1      RandomForest  0.817308   0.799611  0.817308  0.790601
```

Fig 12: Performance metrics of Classification models.

The comparison of classification models forms the fig 12 reveals that **LightGBM** achieves the best performance with an **accuracy of 84.5%** and an **F1 score of 0.841**, demonstrating its ability to effectively capture complex relationships among features such as budget, user reviews, and Facebook likes. Its balanced **precision** (0.843) and **recall** (0.844) suggest that the model not only correctly predicts most categories but also identifies the majority of true instances across all classes.

Gradient Boosting closely follows LightGBM with an **accuracy of 83.9%** and an **F1 score of 0.836**, showing strong predictive capability. While its precision (0.837) and recall (0.839) remain competitive, a slight dip compared to LightGBM indicates occasional difficulty in capturing ambiguous boundary cases, where true positives might be missed.

In contrast, **Random Forest** performs moderately with an **accuracy of 81.7%** and an **F1 score of 0.799**, making it less effective for this task compared to the boosting models. It particularly struggles with mid-range categories such as "Good (6-8)", where overlapping features make classification more challenging. However, Random Forest demonstrates reliable performance in the extreme categories ("Poor" and "Excellent"), where the distinctions are clearer, highlighting its strength in capturing simpler relationships within the data.

The fig 13 shows the confusion matrix of the best performing model Light GBM. The matrix clearly complements its performance metrics showing high no.of true positives and very less no of false positives and false negatives. The “Good” class has more entries due to the use of stratified sampling as the class imbalance prevailed in the test set as well

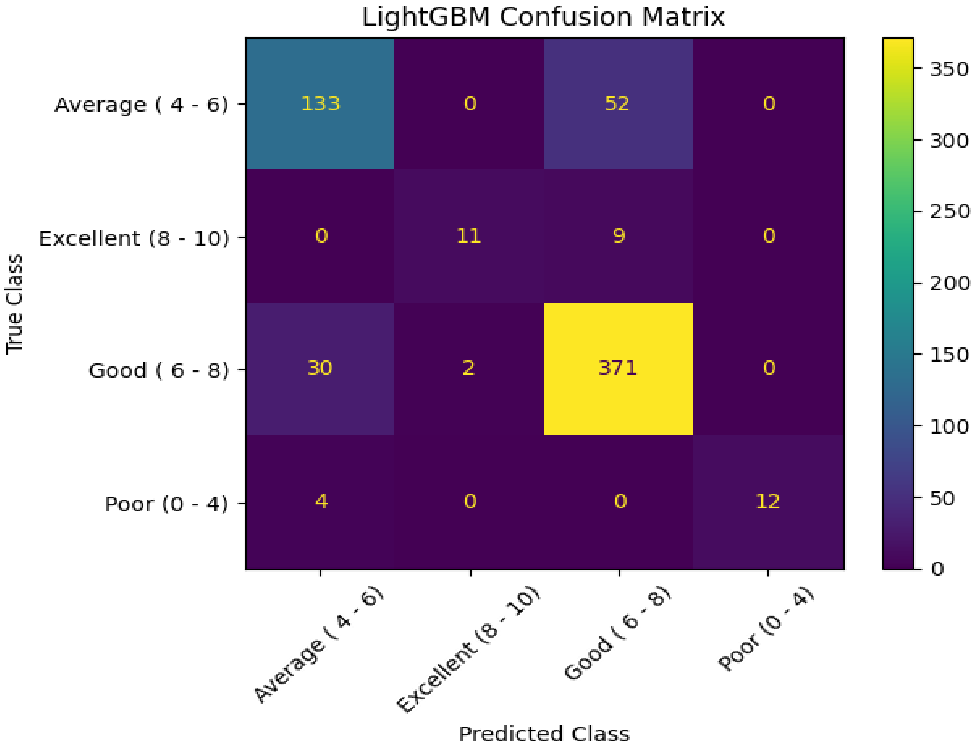


Fig 13: Confusion matrix of LightGBM (Best performing model)

So ultimately the LightGBM models resulted high performance in both the classification and the regression analysis. The classification analysis shows better quality of performance with less error. But we should also consider the impact of the range of each class. With very limited data even regression analysis was predicting significantly better values. Both the regression and classification analysis produced satisfactory performance in predicting the imdb score.

VIII. Conclusion

This study developed a comprehensive predictive framework for anticipating IMDb scores using pre-release movie attributes, addressing key gaps in existing methodologies. By employing advanced regression models such as **LightGBM**, **XGBoost**, and **Gradient Boosting**, alongside classification models for categorizing IMDb scores into meaningful ranges, the approach demonstrated robust performance. The integration of feature engineering techniques, such as **director reputation** and **social media engagement metrics**, played a significant role in enhancing predictive accuracy. LightGBM emerged as the top-performing model, showcasing its ability to capture complex relationships between features. Additionally, the use of GPT models for human-readable recommendations bridges the gap between technical insights and actionable strategies, providing a valuable tool for filmmakers and producers to align production and marketing efforts with audience expectations.

IX. Future Scope

The scope of this study can be further extended by incorporating additional **pre-release features** such as trailer views, social media sentiment trends, and media coverage, which can enhance prediction accuracy. Future work can also focus on predicting other critical metrics like **opening weekend box office revenue**, **lifetime earnings**, and **return on investment (ROI)**, offering broader insights into movie success. Additionally, adapting the framework to other entertainment domains, such as **TV shows**, **web series**, and **video games**, would expand its applicability. A **longitudinal analysis** of trends over time could provide deeper insights into how directors' and actors' performances influence long-term audience reception, further strengthening strategic decision-making capabilities in the entertainment industry.

References:

[1] "Global Movie Production & Distribution - Market Size," IBISWorld, 2024. [Online]. Available: https://www.ibisworld.com/global/market-size/global-movie-production-distribution

[2] "IMDb Developer," IMDb, 2024. [Online]. Available: https://developer.imdb.com/.

[3] E. Bruin, "Movie Recommendation Systems for TMDb," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/code/erikbruin/movie-recommendation-systems-for-tmdb/report.

[4] J. Joseph, et al., "Classifying IMDb Scores into Various Classes," in *Proceedings of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2020. https://ieeexplore.ieee.org/document/8944604/references#references.

[5] "Movie Revenue Prediction," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2405.11651.

[6] A. Jose and S. Harikumar, "Predicting IMDb Movie Ratings Using RoBERTa Embeddings and Neural Networks," in *Responsible Data Science: Lecture Notes in Electrical Engineering*, vol. 940, J. Mathew, G. S. Kumar, and J. M. Jose, Eds. Singapore: Springer, 2022, pp. 231–243. [Online]. Available: https://doi.org/10.1007/978-981-19-4453-6_13.

[7] T. Kumar, S. Kannan, et al., "Predicting IMDb Ratings Using Neural Networks and Pre-Release Attributes," in *Proceedings of IEEE International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9137994.