Predicting Movie Ratings Based on Metadata

Tianren Chen, Sasa Antunovic
* Instructor: Dr. Setarech Rafatirad
[§]University of California, Davis. Davis, CA, USA
tchchen@ucdavis.edu, santunovic@ucdavis.edu

I. Introduction

The rapid growth of digital content, particularly in the film industry, has significantly increased movie production in recent years. This surge has heightened the demand for data-driven tools to predict a movie's reception before release, enabling production companies to create resonant films, streaming platforms to select and recommend content effectively, and viewers to enjoy high-quality movies. This research develops a movie rating predictor that forecasts IMDb ratings using metadata available pre-release, leveraging datasets from the Internet Movie Database (IMDb), Kaggle TMDB Box Office Prediction Dataset, and GitHub Social Media Dataset to capture cinematic, financial, and social media features. To achieve this, we employed Linear Regression, Random Forest Regression, and Artificial Neural Networks, selected for their ability to model diverse relationships within the data [1], [2], [3].

II. Literalter Review

Previous attempts at building IMDb rating predictors [4], [5] have employed similar machine learning techniques to ours for data preprocessing, feature selection, and model training. They followed very similar initial steps for importing and cleaning data from the IMDb database. However, they also imported data from other sources like Kaggle [5] www.the-numbers.com [4]. They did this to be able to use movie gross earnings as an additional feature. Feature selection also slightly differs from our approach as they used features that we found to be negligible like duration, and didn't use features that we found extremely relevant like writers. They also used post-release variables like box office gross income and number of reviews in their models. Tree-based models have emerged as the most effective in these prior efforts [4], [5], achieving prediction accuracies of approximately 75%. These methods were likely chosen for their ability to model non-linear feature interactions and handle mixed data types. However, a notable limitation of these earlier studies is their inclusion of variables like gross income, which are only available after a film's release and thus are not useful for forward-looking predictions or decision-making during film development. In contrast, our approach diverges significantly in terms of both data and methodology. We deliberately exclude features such as box office performance and focus exclusively on attributes that are known or controllable before a movie's release—such as cast, director, genre, production details, and other metadata. This makes our model more relevant for stakeholders looking to estimate potential audience reception in the pre-production or planning stages. Additionally, while we plan to replicate traditional models such as Linear Regression and Random Forest for comparative purposes, our primary focus is on leveraging deep neural networks (DNNs). Prior projects have not explored the use of DNNs in this context, despite their potential to model complex, high-dimensional relationships among input features. We hypothesize that the representational power of deep learning will enable our model to achieve higher predictive accuracy—potentially in the range of 80–90%. Ultimately, by refining the feature set to include only pre-release variables and employing advanced modeling techniques, our goal is to develop a more realistic and actionable IMDb rating predictor that improves upon the benchmarks established in earlier work.

III. Data Description

To develop the movie rating predictor, we utilized three datasets, each providing pre-release features critical to IMDb rating prediction. The IMDb datasets serve as the primary source, supplemented by the Kaggle TMDB Box Office Prediction dataset for financial metrics and a GitHub Social Media Dataset for actor popularity and budget metrics. These features, supported by prior research on movie success factors (4), undergo feature engineering and exploratory data analysis to optimize model performance. The following sections detail each dataset, with key attributes summarized below.

1) ***IMDb Non-Commercial Database****:* The IMDb datasets provide comprehensive metadata on movies, TV shows, and associated personnel. We utilized the following datasets, interconnected via the `tconst` (title identifier) and `nconst` (person identifier) fields for relational analysis. The data point counts reflect estimates as of June 06, 2025.

- *title.basics.tsv.gz* (11,633,024 data points)
  Contains core metadata for titles, including:
  - *titleType*: Type of title (e.g., movie, tvSeries, tvEpisode)
  - *runtimeMinutes*: Duration of the title in minutes (e.g., 120)
  - *genres*: Comma-separated list of up to three genres (e.g., Action,Adventure,Sci-Fi)
  - *startYear*: Release year of the title (e.g., 2020)
- *title.crew.tsv.gz* (11,633,024 data points)
  Links titles to their directors and writers.
  - *directors*, *writers*: Comma-separated list of `nconst` values

- *title.ratings.tsv.gz* (1,564,487 data points)
  Contains IMDb ratings and voting data for titles.
  - *averageRating*: IMDb rating on a 1–10 scale (e.g., 8.5)
  - *numVotes*: Total votes received (e.g., 150,000)
- *name.basics.tsv.gz* (14,378,919 data points)
  Provides information on individuals.
  - *primaryName*: Most commonly used name (e.g., Tom Hanks)
- *title.principals.tsv.gz* (92,343,574 data points)
  Details principal cast and crew for each title.
  - *ordering*: Order of principals (e.g., 1 for lead actor)
  - *category*: Role of the person (e.g., actor, director)

*2) **Kaggle TMDB Box Office Prediction Dataset:*** This dataset provides budget information and other movie metadata, which we merged with IMDb data to incorporate financial features into our model. It includes 4,398 movie entries, each with full budget information and no missing values.

- *budget*: The movie's budget in US dollars.

*3) **GitHub Social Media Dataset:*** The Social Media Dataset offers a variety of movie-related metadata, including director and actor details, genres, ratings, and Facebook like counts for cast members. We selected four features focused on actor popularity.

- *movie_imdb_link*: URL to IMDb (e.g., `https://www.imdb.com/title/tt1234567/`)
- *actor_1_facebook_likes*: Facebook likes for lead actor (e.g., 15,000)
- *actor_2_facebook_likes*: Facebook likes for second actor (e.g., 2,000)
- *actor_3_facebook_likes*: Facebook likes for third actor (e.g., 1,000)

*4) **Dataset Integration:*** To combine the datasets, we used the `tconst` field from the IMDb datasets to link movies across `title.basics.tsv.gz`, `title.crew.tsv.gz`, `title.ratings.tsv.gz`, and `title.principals.tsv.gz`, enabling the extraction of features like genres, directors, and ratings. The `imdb_id` field (or `movie_imdb_link` in the social media dataset) served as the key to merge the Kaggle and social media datasets with the IMDb data, aligning budget and actor popularity features with the corresponding movies for a unified dataset.

To construct a comprehensive dataset, we performed relational joins across all sources. Traditional IMDb metadata was first consolidated using `tconst`, after which external attributes such as `budget` and actor Facebook popularity scores were integrated using `imdb_id`. This unified structure enabled a combined representation of financial, social, and content-based features for robust movie rating prediction.

IV. Feature Engineering

A. *Movie Actor Score vs. IMDb Movie Rating*

(a) This section quantifies the impact of actors on movie ratings using the IMDb database. The actor scores table measures actors' performance and experience through features such as `average_movie_rating`, `total_movies`, `rating_standard_deviation`, `high_rated_movie_count`, `career_span`, and `recent_average_rating` (last 5 years). As shown in Fig 1 Data preprocessing excluded actors with fewer than three movies—650,000 with one and 110,000 with two—to reduce noise. The remaining 222,531 actors included 17,199 with missing `rating_standard_deviations`, which were imputed from others with a `career_span` of 0–5 years. Additionally, 142,180 actors without a `recent_average_rating` had their data filled in with the overall average. Actors with `career_span` over 80 years were also removed.

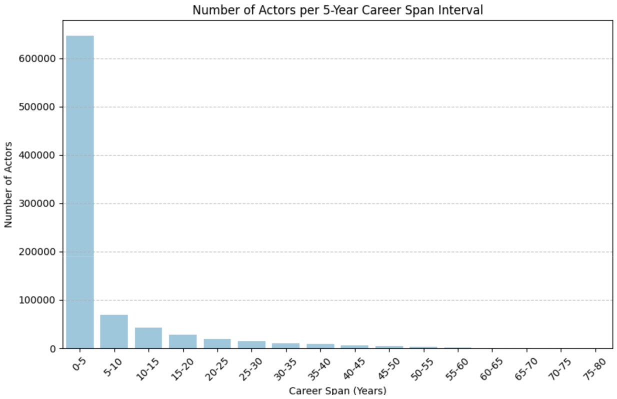

Fig. 1. Actor Score Feature Engineering: This figure visualizes the filtering and preprocessing steps used to derive actor-related features, including exclusion of low-activity actors, imputation of missing values, and removal of outliers.

(b) A Ridge Regression model was trained on movies released between 2020 and 2025 to avoid inflated actor scores from older films. Higher billing order (positions 1–2) was used to identify leading actors, while lower billing order (positions 3–7) represented supporting roles, validating the aggregation of scores for the top seven actors. The model achieved a Mean Squared Error (MSE) of 0.6638, a Root Mean Square Error (RMSE) of 0.8148, and an R2 of 0.7283, successfully explaining 72.83% of the variance in IMDb ratings. Each actor's final score was computed as a weighted sum of selected features and adjusted to ensure non-negative values, enhancing interpretability. To quantify the collective influence of actors on movie ratings, we used the average `actor_score` across up to seven main cast members per movie as a predictive feature for IMDb ratings, labeled as `movie_actor_score`.

B. *Movie Director Quality vs. IMDb Movie Rating*

(c) To investigate the influence of movie directors on IMDb ratings, we analyzed director-related features, including the average movie rating `director_quality`, the total number of movies directed `director_experience`, and the aggregate votes across all movies `director_total_votes`. After cleaning the dataset—excluding 4,094 movies with missing director information and imputing missing release years for 33 movies with the median year (2007)—we processed 371,406 director-movie pairs across 146,578 unique directors.

(d) A Ridge regression model initially incorporated all three features, yielding an R2 score of 0.8699, an MSE of 0.3370, and a test set RMSE of 0.5806. Feature weights revealed director_quality (1.4869) as the dominant factor, with negligible contributions from director_experience (-0.0009) and director_total_votes (-0.0012). A reduced model using only director_quality maintained comparable performance (R2: 0.8699, MSE: 0.3370, RMSE: 0.5805). Consequently, we adopted an average of director_quality for up to three main directors per movie to predict IMDb ratings, labeled as movie_director_quality.

C. Genre Score vs. IMDb Movie Rating

(a) To assess the impact of film genres on IMDb movie ratings, we analyzed average ratings across 28 unique genres, with documentaries averaging the highest at 7.3 and horror films the lowest at 4.9, a 2.4-point difference on the 1–10 scale. Most genres scored between 5.6 and 6.5, with an interquartile rating range of 1.0 to 1.5 points per genre. Genre accounted for 10–20% of rating variance, while 80–90% stemmed from factors like director quality, actor performance, and storytelling. To quantify genre influence, we one-hot encoded genres and calculated average ratings for each. For movies with multiple genres (up to three), we computed a genre score as the weighted sum of average ratings for assigned genres, effectively capturing their cumulative effect on IMDb ratings, labeled as genre_score.

D. Movie Writer Quality vs. IMDb Movie Rating

(a) To investigate the influence of writers on IMDb movie ratings, we analyzed 72,974 unique writers from films released between 2020 and 2025. We focused on two key features: writer_quality (the average rating of a writer's movies) and writer_experience (the number of movies written). A Ridge regression model was developed to predict ratings, achieving an R2 score of 0.8656, Mean Squared Error (MSE) of 0.3223, and Root Mean Squared Error (RMSE) of 0.5677. The feature weights were 1.4457 for writer_quality and -0.0012 for writer_experience. Due to the minimal impact of writer_experience, it was excluded from further analysis. To quantify a movie's cumulative writer impact, we computed a writer score as the average of writer_quality for up to three primary writers per movie, labeled as movie_writer_quality. This method provides a robust and interpretable metric for assessing the contribution of writers in predicting IMDb ratings.

E. Movie Actor Popularity vs. IMDb Movie rating

(a) To examine the relationship between movie actor popularity and IMDb user ratings, I analyzed whether the presence of more popular actors is associated with higher perceived movie quality. Using the Movie_Actor_Popularity feature, calculated as the combined average Facebook likes of the top three billed actors for each film, I compiled a dataset containing 5,010 valid entries with both popularity scores and IMDb ratings. The actor popularity scores ranged from less than 1 to over 218,000, with a mean of approximately 2,967, indicating a highly skewed distribution with a small number of extremely popular actors. Statistical analysis revealed a correlation of 0.012 between Movie_Actor_Popularity and IMDb rating, indicating virtually no linear association between these two variables. Despite this, actor popularity is retained as a feature in modeling, as it may interact with other factors and help improve the accuracy of predictive models.

F. Budget vs. IMDb Movie Rating

(a) As shown in Table I. To analyze the relationship between a movie's production budget and its IMDb user rating, we examined whether financial investment influences perceived movie quality. The initial Kaggle dataset from train.csv contained 3,000 budget entries, but 811 had a budget of $0, indicating unreliable or missing data. After removing these, 2,188 valid entries remained. Merging train.csv with test.csv yielded a total of 7,398 budget entries. To further enhance the dataset, we integrated data from a social media dataset using the imdb_id, recovering 1,593 previously missing budget values. The final dataset consisted of 8,991 budget entries. Although analysis of the merged dataset revealed a correlation of -0.139 between budget and IMDb movie rating—indicating only a weak negative association—budget was retained as a feature in subsequent modeling. This is because budget remains an important descriptor of a film's production scale, and its potential nonlinear effects or interactions with other variables may still enhance model performance. To ensure consistency in financial comparisons across different time periods, all movie budgets were adjusted for inflation using the Consumer Price Index (CPI). We utilized CPI data from the U.S. Bureau of Labor Statistics [6] converting all budget values to their 2025 dollar equivalents. This inflation adjustment mitigates temporal bias and improves the interpretability of financial features.

G. Feature Set Selection

(a) After conducting thorough feature engineering, we identified six key features for training our machine learning models. For the IMDb dataset, which primarily includes movie_actor_score, movie_director_quality, and movie_writer_quality, we chose to focus on movies released between 2020 and 2025. This decision mitigates the risk of inflated feature values, as these features are calculated using data up to the present day, making them less suitable for predicting IMDb ratings for older movies, such as those released 20 years ago. While a time-sensitive approach—recalculating movie_actor_score, movie_director_quality, and movie_writer_quality for each movie based only on data available up to its release date—was considered, it would generate over 3 million data points and significantly

TABLE I
AVERAGE BUDGET PER GENRE (ADJUSTED TO 2025 DOLLARS)

Genre	Avg. Budget (USD)
Animation	$65,012,910.44
Adventure	$63,203,425.01
Sci_Fi	$57,245,524.02
Action	$52,848,120.90
Fantasy	$48,735,279.96
Family	$47,869,763.04
Thriller	$28,614,196.01
Comedy	$28,289,376.57
History	$26,487,368.32
Crime	$25,171,932.30
Mystery	$24,254,080.01
Sport	$24,059,927.67
Biography	$23,846,722.93
Drama	$22,449,168.97
Romance	$20,969,162.45
War	$20,653,775.74
Music	$16,326,453.03
Horror	$16,298,415.35
Musical	$16,117,296.18
Western	$14,616,899.10
Documentary	$4,388,240.25
Film_Noir	$1,184,270.06
News	$350,000.00

increase computational complexity. To balance accuracy and feasibility, we limited our dataset to 53,005 recent movies from the past five years.

(b) To address the disparity between the IMDb dataset (53,005 movies) and the smaller merged dataset of budget and movie popularity data (5,375 movies), we developed two distinct feature sets for model training and evaluation. The first set, codenamed **EXP (Experience)**, encompasses the full recent IMDb dataset and includes `genre_score`, `movie_director_quality`, and `movie_writer_quality`. It excludes actor popularity due to missing data and overlap with `movie_actor_score`, and imputes missing budgets as the average of up to three genre-based averages from Kaggle budget data (see Table II), referred to as `budget_imputed`. The second set, codenamed **POP (Popularity)**, is derived exclusively from the smaller Kaggle and social media merged dataset, incorporating true `budget` and `movie_actor_popularity` data alongside `genre_score`, `movie_director_quality`, and `movie_writer_quality`. By training identical models on these two sets— **EXP** with a larger dataset, `budget_imputed`, and a focus on actor experience, and **POP** with a smaller dataset but accurate budget and popularity data we can effectively compare their performance and robustness.

V. EXPLORATORY DATA ANALYSIS (EDA)

A. *Data Cleaning*

(a) The initial dataset for feature set EXP consisted of 53,005 movies, but only 5,375 of these had non-missing budget values. During the data cleaning process, 10,239 movies (19.3%) were excluded due to missing values in one or more features. Notably, 9,895 movies lacked a movie actor score because actors with fewer than three movie appearances were excluded from the calculations. This filtering affected documentaries and small-scale productions, which often have limited data on actors. The remaining missing features were primarily found in movies with minimal available data, such as documentaries and obscure films that lacked complete records.

(b) About the GitHub Social Media Dataset, the initial dataset included Facebook like counts for the top three billed actors: `actor_1_facebook_likes`, `actor_2_facebook_likes`, and `actor_3_facebook_likes`. However, these values were available for only 5,036 to 5,020 movies due to missing data. Notably, the distribution of Facebook likes was highly skewed, with most actors receiving modest attention, while a few gained exceptionally high counts—up to 640,000 for lead actors, 137,000 for second billed, and 23,000 for third billed. The mean values dropped sharply across roles, from 6,560 for actor 1 to just 645 for actor 3, indicating a steep decline in social media presence beyond the lead role. These missing values were most common in lesser-known films, such as independent productions and documentaries, where actors often have limited or no online presence.

B. *Outlier Analysis*

*a) **Statistics**:* As shown in Table II A comprehensive outlier analysis was conducted on the remaining 42,766 movies to validate the dataset prior to machine learning model training. Summary statistics for the features are shown in %key features are presented below for the reduced dataset that includes valid entries for both `budget` and `movie_actor_popularity`.

TABLE II
ROUNDED SUMMARY STATISTICS FOR SELECTED FEATURES

Feature	Min	Q1	Median	Q3	Max	Outliers
Genre	4.98	5.80	6.03	6.22	7.21	6015
Writer	1.00	5.00	6.00	6.93	10.00	139
Actor	0.57	5.61	6.76	7.53	12.69	422
Budget ($)	0.35M	22.45M	23.90M	28.61M	65.01M	8938
Director	1.00	5.20	6.20	7.10	10.00	214
Popularity	0.67	384.67	727.50	4289.67	218333.33	356

*b) **Movie Actor Score, Director Quality, and Writer Quality**:* The analysis of the distributions revealed that fewer than 1% of the movies were outliers. These outliers represented natural extremes, such as A-list actors exceeding the upper interquartile range (IQR) for `movie_actor_score` or inexperienced writers and directors falling below the lower bound. Since these outliers reflected a small but significant subset of the dataset and were important for robust model training, they were retained.

*c) **Genre Score**:* A total of 6,015 movies (14.1%) were identified as outliers. Among them, 2,376 had `genre_score` below the lower bound (5.17), all belonging to the Documentary genre (score: 4.9753), while 3,639 were above the upper bound (6.86), all belonging to the Horror genre (score: 7.2133). These were preserved as meaningful genre-specific deviations.

*d) **Budget**:* This feature had the highest number of outliers, with 8,938 movies (20.9%) falling outside the interquartile range (IQR) bounds ($13,201,628.41 to $37,861,736.56). This was primarily due to the imputation of 95% of budgets as genre-specific averages (e.g., $4,388,240.25 for documentaries), which narrowed the IQR and made the bounds overly sensitive. The budget distribution was heavily right-skewed, with a long tail of high-budget films. To address this, a log transformation was applied to the `budget_imputed` feature, resulting in an approximately normal distribution suitable for Z-score analysis. Using a threshold of three standard deviations, 2,951 movies (6.9%) were identified as outliers in the log-transformed data. Most of these outliers had the imputed documentary budget of $4,388,240.25, reflecting valid genre specific patterns. These statistically significant outliers were retained for machine learning training to ensure a representative dataset.

*e) **Movie Actor Popularity**:* The outlier analysis revealed a heavily right-skewed distribution, with an interquartile range (IQR) of 3,905.0 and an upper outlier boundary of 10147.1667. A total of 356 outliers were identified above this threshold, with the maximum value reaching 218333.3333. These data highlight the significant presence of extremely popular actors, likely representing top-tier individuals whose popularity substantially exceeds the norm. This trend could be meaningful for IMDb ratings, as highly popular actors might attract larger audiences and potentially boost ratings.

C. Correlation Analysis of Movie Features with IMDb Ratings

This analysis investigates the relationship between 7 movie features – `genre_score`, `movie_writer_quality`, `movie_actor_score`, `movie_director_quality`, `budget` (in millions USD), and `budget_imputed`—and IMDb movie ratings. Pearson correlation coefficients are used to measure the strength and direction of the linear relationship between each feature and the average IMDb rating. The correlation values range from -1 (indicating a perfect negative correlation) to 1 (indicating a perfect positive correlation). These correlations are derived from a dataset of movie features.

*1) **Correlation Results**:* As shown in Fig 2 The bar chart titled ”Correlation of Movie Features with IMDb Ratings” illustrates the Pearson correlation coefficients between the seven movie features and IMDb ratings. We distinguish between the true budget and the imputed budget to justify our choice of two feature sets.

*2) **Interpretation of Correlation Results**:*

*a) **`movie_director_quality`** (0.925): :* This feature exhibits the strongest positive correlation with IMDb ratings. The result indicates that films directed by highly rated or experienced directors tend to receive significantly higher ratings. The vision and expertise of directors play a central role in shaping the narrative, pacing, and emotional resonance of a film, which strongly influences audience and critic reception. `movie_writer_quality` (0.851): A strong positive correlation highlights the importance of high-quality writing—including screenplay structure, dialogue, and narrative complexity—in earning favorable IMDb scores. This suggests that a well-crafted script is a crucial factor in movie success.

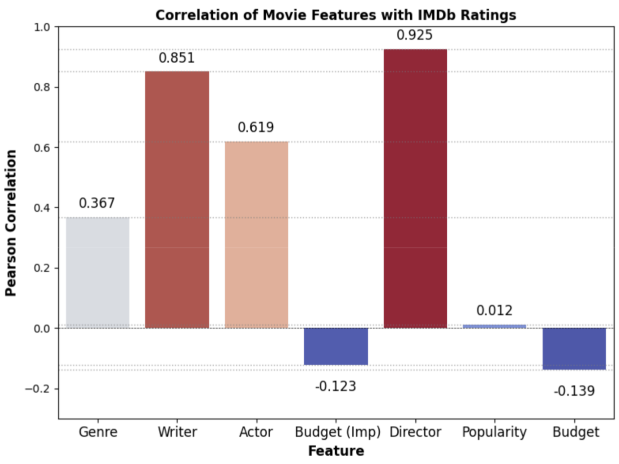

Fig. 2. Correlation of Movie Features with IMDb Ratings. Pearson correlation coefficients for key features such as genre score, actor score, writer quality, director quality, and budget are displayed.

*b) **`movie_actor_score`** (0.619): :* A moderate to strong positive correlation indicates that skilled or high-rated actors contribute to better audience ratings, although the effect is less pro- nounced compared to directors and writers. This underscores the importance of actor performance while acknowledging the greater impact of narrative and direction.

*c) **`genre_score`** (0.367): :* The genre of a movie has a noticeable but relatively moderate impact on its IMDb rating. Some genres, such as drama or thriller, may resonate more positively with audiences, but genre alone is not a dominant predictor of success.

*d) **`budget`** (Millions USD, True, -0.139): :* The correlation between actual budget and IMDb ratings is weak and negative. This suggests that high financial investment does not necessarily translate to higher ratings. It may reflect diminishing returns on expensive productions or the overhyping of under- performing blockbusters.

*e) **`budget_imputed`** (Millions USD, -0.123): :* Similar to true budget values, imputed budgets in the EXP dataset show a weak negative correlation. This supports the decision to evaluate both real and imputed budget data across separate feature sets.

*f) **`movie_actor_popularity`** (0.012): :* The weakest correlation among all features analyzed, this result shows that actor popularity on social media has virtually no linear relationship with IMDb ratings. While popularity may correlate with visibility or marketing, it does not consistently predict audience evaluation.

3) Conclusion: (a) The correlation analysis justifies the inclusion of movie_actor_score in both the EXP and POP feature sets due to its substantially higher correlation (0.619) compared to movie_actor_popularity (0.012). Although popularity provides contextual social data, actor quality proves to be a far more reliable predictor. (b) The difference in correlation between budget_imputed values in EXP (-0.123) and true budget values in POP (-0.139) also supports the dual-feature set strategy. While the difference is small, it demonstrates that the smaller POP dataset captures slightly more accurate financial influence on IMDb ratings. This dual approach allows for performance comparison between a large, comprehensive dataset (EXP) and a smaller, high-fidelity dataset (POP), enabling a trade-off analysis between data size and quality for predictive modeling.

D. Feature-to-Feature Correlation Analysis

Feature-to-feature correlation analysis was performed to examine relationships between different features and evaluate our choice to use two different feature sets, EXP and POP. Feature-to-feature correlations are reported in heatmap Fig 3

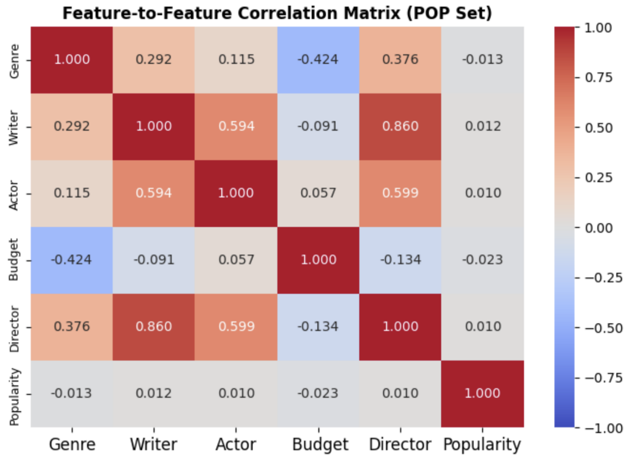

Fig. 3. Feature-to-Feature Correlation Heatmap. Pearson correlation coefficients between features in the POP dataset reveal relationships that inform feature independence and model construction.

a) ***Independence of Actor Popularity :*** movie_actor_popularity correlates weakly with other POP features: -0.013 (Genre), 0.012 (Writer), 0.010 (Actor), -0.023 (Budget), and 0.010 (Director), with an average correlation of 0.0008. This near-zero average highlights its independence, indicating that it captures a distinct aspect—namely, actor fame—without introducing redundancy. Despite its low correlation (0.012) with IMDb ratings, this uniqueness justifies its inclusion in the POP feature set.

b) ***High Director–Writer Correlation:*** The 0.860 correlation between movie_director_quality and movie_writer_quality in POP likely stems from directors and writers being the same person or frequent collaborators. This high correlation risks multi-collinearity, so we'll monitor their impact in EXP and POP models, potentially combining them or using regularization.

c) ***Budget–Genre Correlation:*** The -0.424 correlation between budget and genre_score in POP supports imputing budgets in EXP using genre averages. Stronger than budget (-0.139) or budget_imputed (-0.123) correlations with IMDb ratings, it validates our dual EXP and POP approach to balance dataset size and data accuracy.

d) ***Weak Correlation Between Actor Popularity and Score:*** The weak Pearson correlation of 0.010 between movie_actor_popularity (top_3_actor_scores) and movie_actor_score in the POP dataset may seem counterintuitive, as both relate to actors, but is explained by their construction. movie_actor_score averages the quality of seven actors, capturing a broader performance metric, while movie_actor_popularity reflects the fame of only three actors. Additionally, since popularity data relies on actor names, it may include actors not in the movie_actor_score calculation, leading to mismatches that dilute their correlation. This independence supports including both in POP to capture distinct aspects, with movie_actor_score (0.619 correlation with IMDb ratings) driving predictions and movie_actor_popularity (0.012) offering unique contextual insights.

VI. Methodology

A. Model Selection

(a) To predict IMDb movie ratings, we evaluated three machine learning models—Linear Regression, Artificial Neural Network (ANN), and Random Forest (RF)—chosen for their complementary strengths in prior studies. Linear Regression was selected for its interpretability, as shown by [7] and [8] in predicting ratings using movie metadata. ANN was included for its ability to model non-linear relationships, validated by. [9], while Random Forest was chosen for its robustness to overfitting, as supported by [10]. These models enable a comprehensive comparison of linear, non-linear, and ensemble approaches.

B. Model definition

a) Linear regression: a statistical method that models a linear relationship between a dependent variable and one or more independent variables, where prediction is a weighted sum of the input features plus an intercept, optimized by minimizing the mean squared error (MSE).

b) An Artificial Neural Network (ANN): a machine learning technique that learns complex patterns through interconnected layers of nodes (neurons) that process input features through weighted connections, applying activation functions to capture non-linear relationships.

c) Random Forest (RF): is a machine learning algorithm that creates multiple independent decision trees based on a random subset of data and features, and predicts by averaging the predictions of all trees.

C. Evaluation Metrics

a) Evaluation Performance: Model performance was assessed using mean squared error (MSE), root mean squared error (RMSE), and R-squared (R²), standard metrics for regression tasks [7], [8]. MSE measures the average squared difference between predicted and actual ratings, emphasizing larger errors and providing a precise measure of prediction accuracy. RMSE, the square root of MSE, is expressed in the original rating scale, making it intuitive for comparing model performance. R² indicates the variance in the ratings explained by the model. These metrics were selected to ensure a comprehensive evaluation of model accuracy, error magnitude, and goodness of fit.

b) Mathmatic Formular:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

D. Shared methodology

(a) For each of the three models created, features were standardized using `StandardScaler` to ensure comparable scales. The methodology leveraged Python libraries: `pandas` for data handling, `numpy` for computations, `scikit-learn` for preprocessing (`StandardScaler`, `train_test_split`, `GridSearchCV`), modeling (`LinearRegression`, `RandomForestRegressor`), and evaluation (`mean_squared_error`, `r2_score`), `tensorflow` and `keras` for ANN implementation (`Sequential`, `Dense`, `SGD`), and `scikeras` for Keras integration. Model performance was assessed using R^2, MSE, and RMSE, with results saved as CSV files.

E. Models Implementations

a) Linear Regression: A linear regression model was trained on standardized features using 5-fold cross-validation, with shuffling to reduce bias and mitigate overfitting, as is standard in regression tasks [7]. The model was trained on each fold, with performance measured using R^2, MSE, and RMSE. Average feature coefficients were calculated across folds to interpret feature importance.

b) Artificial Neural Network (ANN): An ANN was developed with a single hidden layer, standardized features, and hyperparameter tuning (neurons, activation, momentum, batch size) was performed via 3-fold cross-validation with grid search, using early stopping to prevent overfitting. The hyperparameter grid for the Artificial Neural Network (ANN) model was chosen to optimize performance while balancing computational efficiency. Neuron counts (32, 64, 128) were selected to test a range of network capacities, from simple to complex. Activation functions (tanh, ReLU) were included to address the handling of non-linear patterns, as supported by prior studies [9]. Momentum values (0.6, 0.9) and batch sizes (32, 64) were selected to investigate the trade-offs between training stability and convergence speed. The best model was trained on a randomly selected 80/20 train-test split and evaluated using R^2, MSE, and RMSE.

c) Random Forest (RF): To predict IMDb movie ratings, we implemented a Random Forest Regressor. Features were standardized, and an 80/20 train-test split was applied. Hyperparameter tuning (n_estimators, max__depth, and min_samples_split) was performed using 3-fold cross-validation with a grid search to optimize performance. The hyperparameter grid was selected to find a combination of hyperparameters that balances model complexity and performance. The best model was evaluated using R^2, MSE, and RMSE. The number of trees (n_estimators: 100, 200) was chosen to achieve accurate predictions with acceptable computational cost. Maximum tree depth (max_depth: 10, 20, None) and minimum samples per split (min_samples_split: 2, 5) were included to explore the trade-offs between overfitting and generalization.

d) Pipeline Graphic: A modular pipeline that processes IMDb, Kaggle, and social data through EDA and feature selection to build, evaluate, and compare EXP and POP models for final deployment.

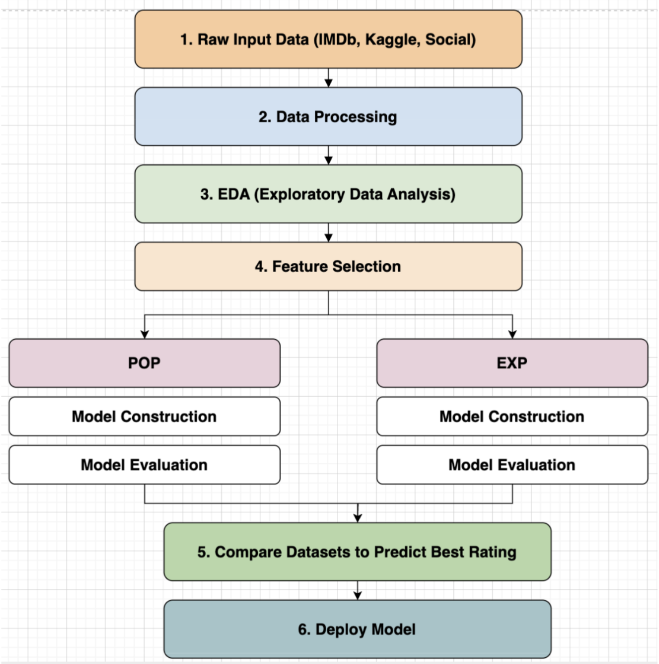

Fig. 4. End-to-End Machine Learning Pipeline. This diagram illustrates the full process from feature preprocessing and model selection to training, tuning, and evaluation.

VII. Experimental Results and Evaluation

A. Dataset with traditional IMDb (Baseline)

1) Linear Regression:

a) Model Performance Analysis: The linear regression model, evaluated using 5-fold cross-validation on the EXP dataset, exhibits robust performance in predicting IMDb movie ratings. As shown in Table III, with an average R² score of 0.874 (±0.003), indicating that the features explain 87.4% of the variance in ratings. The low standard deviations in R² (0.003), MSE (0.01), and RMSE (0.009) demonstrate consistent performance, with an average RMSE of 0.568 (±0.009), indicating that predictions deviate by approximately 0.57 points on the IMDb scale.

TABLE III
LINEAR REGRESSION MODEL PERFORMANCE ON EXP DATASET (5-FOLD CROSS-VALIDATION)

Metric	R^2	MSE	RMSE
Mean	0.874	0.322	0.568
Standard Deviation	0.003	0.010	0.009

b) Feature Coefficients: As shown in Table IV, feature coefficients emphasize `movie_director_quality` (1.097) as the primary predictor, followed by `movie_writer_quality` (0.309) and `movie_actor_score` (0.149), aligning with their strong correlations (e.g., 0.619 for `movie_actor_score`). `genre_score` (0.073) has a slight positive effect, while `budget_imputed` (-0.016) shows a minor negative impact, consistent with its weak correlation (-0.123). The high `movie_director_quality` coefficient, along with its 0.860 correlation with `movie_writer_quality`, suggests the presence of multicollinearity, potentially amplified by the imputed values in the EXP dataset. These results validate the effectiveness of the EXP dataset (53,005 movies) in leveraging key features for accurate IMDb rating predictions.

TABLE IV
LINEAR REGRESSION COEFFICIENTS ON EXP DATASET

Feature	Coefficient
Genre Score	0.073
Writer Quality	0.309
Actor Score	0.149
Budget (Millions USD)	-0.016
Director Quality	1.097

2) *Artificial Neural Network (ANN):*

a) Model Performance Analysis: As shown in Table V, the Artificial Neural Network (ANN) model, optimized for the EXP dataset (53,005 movies) using 3-fold cross-validation, predicted IMDb ratings with an average MSE of 0.275 and RMSE of 0.524 across folds, indicating that predictions deviated by approximately 0.52 rating points. The final model, trained on an 80/20 train-test split, achieved an R^2 score of 0.888, explaining 88.8% of the variance in ratings, with an MSE of 0.284 and RMSE of 0.533. These results underscore the ANN's ability to leverage EXP features, such as `movie_actor_score` (correlation 0.619) and `budget_imputed` (correlation -0.123), to capture complex relationships and deliver robust IMDb rating predictions.

TABLE V
ANN MODEL PERFORMANCE ON EXP DATASET

Model with Best Hyperparameters (3-Fold Average)	
Average MSE	0.275
Average RMSE	0.524
Final Model Performance (80/20 Train-Test Split)	
R^2 Score	0.888
MSE	0.284
RMSE	0.533

b) Hyperparameter Performance Analysis: As shown in Table VI: Hyperparameter analysis for the ANN on the EXP dataset showed that ReLU activation (avg. RMSE = 0.539) outperformed Tanh (0.577), leveraging its non-linear properties. A momentum of 0.9 (RMSE = 0.548) surpassed 0.6 (0.568), enhancing convergence. A neuron count of 32 (RMSE = 0.555) yielded better results than 64 (RMSE = 0.558) or 128 (RMSE = 0.562), suggesting that a simpler model avoids overfitting. Batch size of 32 (RMSE = 0.554) slightly outperformed 64 (0.563). With low standard deviations (e.g., 0.010 for ReLU), these findings validate the optimal configuration (32 neurons, ReLU, 0.9 momentum, 32 batch size, RMSE = 0.524) for effectively modeling feature interactions in the EXP dataset.

TABLE VI
ANN HYPERPARAMETER PERFORMANCE ON EXP DATASET

Hyperparameter	Val	Avg MSE	Std MSE	Avg RMSE	Std RMSE
Activation	ReLU	0.291	0.010	0.539	0.010
Activation	Tanh	0.334	0.019	0.577	0.016
Momentum	0.6	0.323	0.025	0.568	0.022
Momentum	0.9	0.301	0.024	0.548	0.021
Neuron Count	32	0.309	0.026	0.555	0.023
Neuron Count	64	0.311	0.028	0.558	0.025
Neuron Count	128	0.316	0.029	0.562	0.026
Batch Size	32	0.307	0.024	0.554	0.022
Batch Size	64	0.317	0.029	0.563	0.025

3) *Random Forest (RF):*

a) Model Performance Analysis: As shown in Table VII: The Random Forest model's feature importance analysis reveals that `movie_director_quality` dominates the predictive signal, with an importance score of 0.876. This indicates that the perceived quality of the director plays a decisive role in determining a film's IMDb rating. In contrast, features such as `movie_writer_quality` (0.077) and `movie_actor_score` (0.030) contribute modestly to the model's predictions. The `movie_actor_popularity` (0.007), `genre_score` (0.005), and `budget_imputed` (0.005) show minimal influence, suggesting that either their signal is weak or their effect is captured indirectly through more influential features. This hierarchy implies that while social metadata can add incremental predictive value, traditional quality-related attributes (especially director-related) remain the most reliable indicators in ensemble models like Random Forest.

TABLE VII
RANDOM FOREST MODEL PERFORMANCE ON EXP DATASET (3-FOLD CROSS-VALIDATION)

Model (Fold)	R^2 Score	MSE	RMSE
Random Forest (3-fold avg.)	0.909	0.230	0.480

b) Best Hyperparameters Interpretation: As shown in Table XV, the 5-fold grid search selected `n_estimators` = 100, `max_depth` = 10, and `min_samples_split` = 5. With 100 trees, the model strikes a balance between computational efficiency and variance reduction. A maximum depth of 10 helps prevent overfitting, improving the model's generalization to unseen movies in the EXP dataset. Setting `min_samples_split` to 5 avoids overly specific splits, thereby enhancing model robustness. These hyperparameter choices optimize the model's ability to capture complex relationships among EXP dataset features, supporting the high R^2 and low RMSE performance.

TABLE VIII
BEST HYPERPARAMETERS FOR RANDOM FOREST MODEL

Parameter	Value
`n_estimators`	100
`max_depth`	10
`min_samples_split`	5

c) Feature Importance Interpretation: As shown in Table IX: The Random Forest model's feature importance analysis reveals that Director Quality dominates the predictive signal, with an importance score of 0.876. This indicates that the perceived quality of the director plays a decisive role in determining a film's IMDb rating. In contrast, features such as Writer Quality (0.077) and Actor Score (0.030) contribute modestly to the model's predictions. The Top 3 Actor Social Media Score (0.007), Genre Score (0.005), and Budget (0.005) show minimal influence, suggesting that either their signal is weak or their effect is captured indirectly through more influential features. This hierarchy implies that while social metadata can add incremental predictive value, traditional quality-related attributes (especially director-related) remain the most reliable indicators in ensemble models like Random Forest.

TABLE IX
FEATURE IMPORTANCES FOR RANDOM FOREST MODEL

Feature	Importance
Genre Score	0.0030
Writer Quality	0.1120
Actor Score	0.0220
Budget (Millions USD)	0.0030
Director Quality	0.8600

Model Comparison

As shown in Table 5, this evaluation highlights the superiority of the Random Forest algorithm in predicting IMDb ratings for the EXP dataset, which includes 53,005 movies. The Random Forest model achieves an R2 score of 0.909 and a Root Mean Square Error (RMSE) of 0.487, indicating an average prediction error of about 0.49 points on a 1–10 scale. Using an ensemble of 100 trees, the model effectively captures non-linear dynamics and emphasizes the importance of the feature `movie_director_quality`, which has an importance score of 0.86. In contrast, Linear Regression—identified as the least effective model—reports an R2 of 0.874 and an RMSE of 0.568, corresponding to an average error of approximately 0.57 points. Its linear assumptions prove insufficient for modeling the complexity of the EXP dataset. Additionally, multicollinearity between `movie_director_quality` (coefficient: 1.097) and `movie_writer_quality` further limits its ability to capture nuanced influences on ratings.

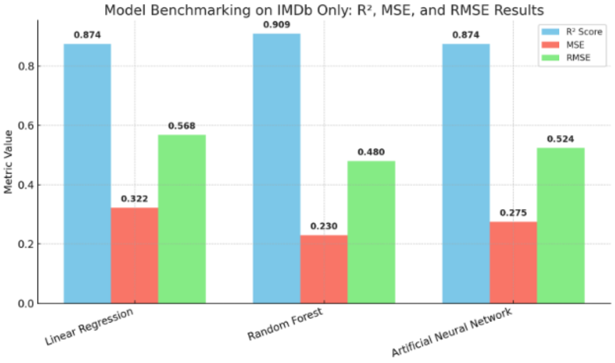

Fig. 5. Model Evaluation Results on IMDb Dataset. This visualization compares R², MSE, and RMSE for different models trained on the EXP dataset.

B. *Dataset with traditional + Social Metadat: POP*

1) *Linear Regression (LR):*

a) Model Performance Analysis: As shown in Table X: The Linear Regression model, trained with traditional and social metadata features, achieved a 5-fold average R² score of 0.882, a Mean Squared Error (MSE) of 0.306, and a Root Mean Squared Error (RMSE) of 0.553. The inclusion of social meta-data features resulted in a modest performance improvement.

TABLE X
LINEAR REGRESSION MODEL PERFORMANCE ON EXP DATASET (5-FOLD CROSS-VALIDATION)

Metric	R^2	MSE	RMSE
Mean	0.882	0.306	0.553
Standard Deviation	0.006	0.013	0.012

b) Feature Coefficients: As shown in Table XI: From this linear Regression Coefficient table shows: `movie_director_quality` shows the strongest positive influence (1.178), followed by `movie_writer_quality` and `movie_actor_score`. Interestingly, the `budget` has a slightly negative coefficient, suggesting diminishing returns for expensive films. Social media influence is small but non-zero.

2) *Artificial Neural Network (ANN):*

TABLE XI
LINEAR REGRESSION COEFFICIENTS ON EXP DATASET

Feature	Coefficient
Genre Score	0.061
Writer Quality	0.251
Actor Score	0.141
Budget (Millions USD)	-0.026
Director Quality	1.178
Movie Actor Popularity	0.003

a) Model Performance Analysis: As shown in Table XII: Since, our output presents the evaluation results of the Artificial Neural Network (ANN) model. With 3-fold cross-validation, the ANN achieved an average R² score of 0.899, indicating it explains about 90 percent of the variance in IMDb ratings. The mean squared error was 0.254, and the root mean squared error was 0.504, both reflecting strong predictive accuracy. These results were consistent when tested on a final 80/20 train-test split, suggesting that the model generalizes well to new, unseen data. Overall, the ANN demonstrates high accuracy and reliability for movie rating prediction in this analysis.

TABLE XII
ANN MODEL PERFORMANCE ON EXP DATASET

Model with Best Hyperparameters (3-Fold Average)	
Average MSE	0.258
Average RMSE	0.508
Final Model Performance (80/20 Train-Test Split)	
R^2 Score	0.899
MSE	0.254
RMSE	0.504

b) Hyperparameter Performance Analysis: As shown in Table XIII: Since, we using the grid search tested various ANN hyperparameter settings by adjusting neurons, activation function, momentum, and batch size. The best combination used 64 neurons, relu activation, 0.9 momentum, and a batch size of 32, which gave the lowest cross-validated MSE of 0.258 and RMSE of 0.508. On the final test split by 80 for training and 20 for testing (80/20), this optimized ANN achieved R² of 0.899, MSE of 0.254, and RMSE of 0.504, demonstrating strong predictive performance.

TABLE XIII
TOP 5 BEST-PERFORMING ANN HYPERPARAMETER COMBINATIONS USING 3-FOLD CROSS-VALIDATION

Activation	Momentum	Batch Size	Mean MSE	Mean RMSE
ReLU	0.9	32	0.258	0.508
ReLU	0.9	32	0.259	0.509
ReLU	0.9	64	0.259	0.509
ReLU	0.9	64	0.260	0.510
Tanh	0.9	32	0.260	0.510

3) Random Forest (RF):

a) Model Performance Analysis: As shown in Table XIV: The Random Forest model's feature importance analysis reveals that Director Quality dominates the predictive signal, with an importance score of 0.876. This indicates that the perceived quality of the director plays a decisive role in determining a film's IMDb rating. In contrast, features such as `movie_qriter_quality` (0.077) and `movie_actor_score` (0.030) contribute modestly to the model's predictions. The `movie_actor_popularity` (0.007), `genre_score` (0.005), and `budget` (0.005) show minimal influence, suggesting that either their signal is weak or their effect is captured indirectly through more influential features. This hierarchy implies that while social metadata can add incremental predictive value, traditional quality-related attributes (especially director-related) remain the most reliable indicators in ensemble models like Random Forest.

TABLE XIV
RANDOM FOREST MODEL PERFORMANCE ON EXP DATASET (3-FOLD CROSS-VALIDATION)

Model (Fold)	R^2 Score	MSE	RMSE
Random Forest (3-fold avg.)	0.905	0.240	0.490

b) Best Hyperparameters Interpretation: As shown in Table XV: The final hyperparameters for the Random Forest model—100 estimators, a maximum depth of 10, and a minimum sample split by 5. Those outputs were identified through grid search using 3-fold cross-validation. This systematic search across predefined parameter combinations aimed to optimize model performance while avoiding overfitting. The selected configuration achieved the lowest cross-validated MSE and RMSE values, indicating a well-balanced trade-off between model flexibility and generalization. The grid search ensured that the model was not arbitrarily tuned, but rather adjusted based on empirical performance over validation folds.

TABLE XV
BEST HYPERPARAMETERS FOR RANDOM FOREST MODEL

Parameter	Value
n_estimators	100
max_depth	10
min_samples_split	5

c) Feature Importance Interpretation: As shown in Table XVI: The Random Forest model's feature importance analysis reveals that Director Quality dominates the predictive signal, with an importance score of 0.876. This indicates that the perceived quality of the director plays a decisive role in determining a film's IMDb rating. In contrast, features such as Writer Quality (0.077) and Actor Score (0.030) contribute modestly to the model's predictions. The Top 3 Actor Social Media Score (0.007), Genre Score (0.005), and Budget (0.005) show minimal influence, suggesting that either their signal is weak or their effect is captured indirectly through more influential features. This hierarchy implies that while social metadata can add incremental predictive value, traditional quality-related attributes (especially director-related) remain the most reliable indicators in ensemble models like Random Forest.

TABLE XVI
FEATURE IMPORTANCES FOR RANDOM FOREST MODEL

Feature	Importance
Genre Score	0.0052
Writer Quality	0.0773
Actor Score	0.0297
Budget (Millions USD)	0.0046
Director Quality	0.8758
Movie Actor Popularity	0.0074

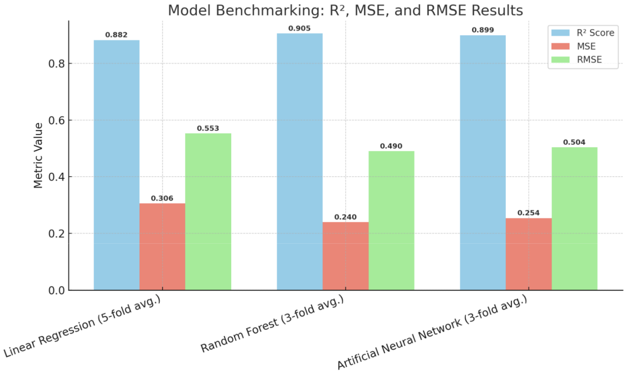

Fig. 6. Model Evaluation Results on IMDb Dataset. This visualization compares R^2, MSE, and RMSE for different models trained on the POP dataset.

Model Comparison

C. *Model Benchmarking: IMDb vs. IMDb + Social Features*

Summary: As shown in Table XVII, The evaluation shows that, when using only IMDb features, Random Forest achieved the best performance with an R^2 score of 0.909, MSE of 0.230, and RMSE of 0.480. Both Linear Regression and Neural Network models performed similarly, each with an R^2 of 0.874, but Neural Network had lower error values. When social media features were added, model performances changed: Random Forest's R^2 decreased slightly to 0.905, while both Linear Regression and Neural Network improved to 0.882 and 0.905, respectively. Notably, the Neural Network saw the most improvement with social features, matching Random Forest's results and outperforming Linear Regression in both R^2 and error metrics. Overall, adding social meta-data led to modest gains for Linear Regression and significant gains for Neural Network, while Random Forest remained the top or joint-top performer across datasets.

TABLE XVII
MODEL PERFORMANCE COMPARISON ON IMDB ONLY VS. IMDB + SOCIAL DATASETS

Dataset	Model	R^2 Score	MSE	RMSE
IMDb Only	Linear Regression	0.874	0.322	0.568
IMDb Only	Random Forest	0.909	0.230	0.480
IMDb Only	Neural Network	0.874	0.275	0.524
IMDb + Social	Linear Regression	0.882	0.297	0.545
IMDb + Social	Random Forest	0.905	0.240	0.490
IMDb + Social	Neural Network	0.905	0.240	0.490

VIII. CONCLUSION

(a) This research demonstrated that the Random Forest algorithm outperforms other models in predicting IMDb ratings across both the EXP and POP datasets. The quality of the movie director emerged as a key predictor in generating accurate pre-release forecasts. The larger EXP dataset enhanced the model's ability to generalize, while the smaller, more focused POP dataset helped validate the consistency and reliability of the selected features. This project offers production companies insights into prioritizing impactful directors and assists streaming platforms in curating engaging content. However, limitations such as imputed budget values in the EXP dataset and restricted actor metrics constrained the model's overall performance.

(b) Future research could investigate gradient boosting techniques to improve model accuracy. Additionally, incorporating time-sensitive computations for `movie_director_quality`, `movie_actor_score`, and `movie_writer_quality` could expand the dataset's depth. Enhancing the `movie_actor_score` by distinguishing the influence of lead and supporting actors, along with using complete and verified budget data, would further increase prediction accuracy.

(c) This project provided our team with valuable hands-on experience in applying machine learning concepts learned in class. The collaboration and persistence invested throughout the quarter strengthened our enthusiasm for pursuing careers in software engineering. We are grateful for this opportunity and look forward to expanding our knowledge in artificial intelligence and machine learning, with the goal of contributing meaningfully to society and making a positive global impact.

REFERENCES

[1] Kaggle. (2025) Tmdb box office prediction. Accessed: 2025-05-05. [Online]. Available: https://www.kaggle.com/c/tmdb-box-office-prediction/data

[2] IMDb. (2025) Imdb non-commercial datasets. Accessed: 2025-05-19. [Online]. Available: https://developer.imdb.com/non-commercial-datasets/

[3] GitHub Repository. (2025) Facebook and imdb metadata collection. Accessed: 2025-05-05. [Online]. Available: https://github.com/sundeepblue/movie_rating_prediction

[4] J. Kunzler. (2021) Choosing the best regression model – imdb movie rating prediction. Accessed June 8, 2025. [Online]. Available: https://medium.com/@jingkunzler211/choosing-the-best-regression-model-imdb-movie-rating-prediction-3298fb11b6d

[5] S. Anand. (2020) Imdb score prediction for movies. Accessed June 8, 2025. [Online]. Available: https://www.kaggle.com/code/saurav9786/imdb-score-prediction-for-movies

[6] U.S. Bureau of Labor Statistics. (2025) Consumer price index (cpi). Accessed: 2025-05-05. [Online]. Available: https://www.bls.gov/cpi/

[7] P.-Y. Hsu, Y.-H. Shen, and X.-A. Xie, "Predicting movies user ratings with imdb attributes," in *Lecture Notes in Computer Science*, 2014, pp. 444–453.

[8] H. Chamani, Z. S. Zadeh, and B. Bahrak, "An overview of regression methods in early prediction of movie ratings," in *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, Oct 2021, pp. 1–6.

[9] W. R. Bristi, Z. Zaman, and N. Sultana, "Predicting imdb rating of movies by machine learning techniques," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul 2019.

[10] S. M. Abidi, Y. Xu, J. Ni, X. Wang, and W. Zhang, "Popularity prediction of movies: From statistical modeling to machine learning techniques," *Multimedia Tools and Applications*, vol. 79, no. 47–48, pp. 35 583–35 617, Jan 2020.

[11] H. Almeida, D. Guedes, W. Meira, and M. Zaki, "Is there a best quality metric for graph clusters," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2011.

[12] W. R. Bristi, Z. Zaman, and N. Sultana, "Predicting imdb rating of movies by machine learning techniques," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019.

[13] J. Kunzler. (2021) Choosing the best regression model - imdb movie rating prediction. Accessed: 2025-05-05. [Online]. Available: https://medium.com/@jingkunzler211/choosing-the-best-regression-model-imdb-movie-rating-prediction-3298fb11b6d

[14] Saurav9786. (2020) Imdb score prediction for movies. Accessed: 2025-05-05. [Online]. Available: https://www.kaggle.com/code/saurav9786/imdb-score-prediction-for-movies

[15] K. Purohit, "Separation of data cleansing concept from eda," https://www.researchgate.net/publication/353162033_Separation_of_Data_Cleansing_Concept_from_EDA, 2021, accessed: 2025-05-19.