SABANCI UNIVERSITY

DSA-210 FINAL REPORT

IMDb vs. Rotten Tomatoes: Predicting IMDb Ratings with Machine Learning

Şeyma Şahin 28033

PROJECT TITLE: SUPERVISED BY: Selim Balcısoy – Kerem Aydın

30.05.2025

ABSTRACT

- The following report investigates the relationship between movie ratings from IMDb and Rotten Tomatoes. It explores whether public metadata such as Rotten Tomatoes scores, Metacritic ratings, genre, runtime, and Oscar wins can be used to predict a movie's IMDb score.
- For this purpose, a dataset of 250 popular movies was analyzed. By employing data cleaning, feature engineering, exploratory data analysis (EDA), and supervised machine learning techniques (specifically, Random Forest Regression), this work sought to identify key predictive features and evaluate the model's performance in forecasting IMDb ratings.
- **Parameters/Features in the Report:**
  - **Title:** The name of the movie.
  - **Year:** The release year of the movie.
  - **Decade:** The decade of the movie's release (derived from Year).
  - **Genre (MainGenre_Label):** The primary genre of the movie.
  - **Runtime:** The duration of the movie in minutes.
  - **Awards (Won_Oscar):** A binary indicator of whether the movie won an Oscar.
  - **IMDb Rating:** The movie's rating on IMDb (0-10 scale) - *This is the target variable for prediction.*
  - **Rotten Tomatoes (RT_Rating):** The movie's score on Rotten Tomatoes (0-100 scale).
  - **Metacritic:** The movie's score on Metacritic (0-100 scale).
  - **MainGenre_ForModel_...:** One-hot encoded genre features used in the model.

1. INTRODUCTION

In today’s digital era, film ratings play a crucial role in how audiences discover and judge movies. Among the many platforms available, IMDb and Rotten Tomatoes are two of the most widely referenced sources for movie evaluations. While IMDb reflects public sentiment through user ratings, Rotten Tomatoes typically incorporates both critic and audience scores. This project aims to explore the relationship between these two platforms and investigate whether IMDb ratings can be predicted using metadata and alternative ratings such as those from Rotten Tomatoes and Metacritic. It also questions the consistency between critical and public evaluations of films. To achieve this, we perform data cleaning, feature transformation, exploratory data analysis, and ultimately develop a supervised machine learning model to predict IMDb scores. By analyzing feature importance, we further understand which factors (e.g., genre, runtime, awards) most significantly influence public ratings. The findings of this project provide insight into how metadata and third-party scores contribute to audience perception and whether aggregated metadata can act as a proxy for IMDb’s crowd-based evaluations.

2.METHODOLOGY

This study utilized a publicly available dataset (`movies-250.json`), which contains metadata for 250 popular films. The dataset was structured in JSON format and included nested fields such as `Ratings`, where Rotten Tomatoes and Metacritic scores were stored alongside other rating sources.

- **Data Preprocessing and Feature Engineering**
    - The data was first loaded into a Pandas DataFrame. A custom Python function was then used to extract numerical values for `RT_Rating` and `Metacritic` from the nested `Ratings` field. The IMDb score (`imdbRating`) was converted to a numeric format to ensure consistency.
    - The `Runtime` field, originally formatted as strings like “123 min”, was cleaned by removing the unit and converting the values into numeric. The release `Year` was also converted to numeric, and a new feature `Decade` was derived from it to analyze temporal trends.
    - To enrich the dataset, a binary feature called `Won_Oscar` was created based on the `Awards` field, marking whether a film had won at least one Oscar. The primary genre for each film was extracted into `MainGenre_Label` for visualization purposes, while a separate column `MainGenre_ForModel` was created and one-hot encoded to be used as machine learning input features.
    - After the preprocessing steps, missing values in the feature matrix `X` were handled by applying `dropna()` to ensure clean model training input.

- **Exploratory Data Analysis (EDA)**
    - To understand the distributions and relationships among the variables, various EDA techniques were employed using Matplotlib and Seaborn libraries. Histograms were plotted for both `imdbRating` and `RT_Rating` to inspect their distributions. A scatter plot of `imdbRating` vs. `RT_Rating` was created to visualize their correlation, and the Pearson correlation coefficient was calculated to quantify this relationship.
    - Box plots were used to analyze the distribution of IMDb scores across different genres, as defined by `MainGenre_Label`. Additionally, a bar chart showing average IMDb rating by decade was created to explore temporal trends in film ratings.
- **Machine Learning: Predicting IMDb Rating**
    - A supervised machine learning approach was implemented to predict IMDb ratings, treating it as a regression task. The feature set `X` included the following variables: `RT_Rating`, `Metacritic`, `Runtime`, `Decade`, `Won_Oscar`, and one-hot encoded genre columns derived

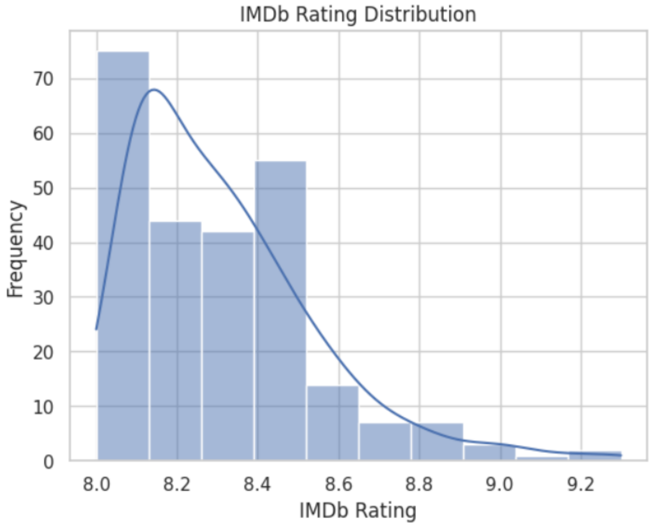
from `MainGenre_ForModel`. The target variable `y` was defined as `imdbRating`.

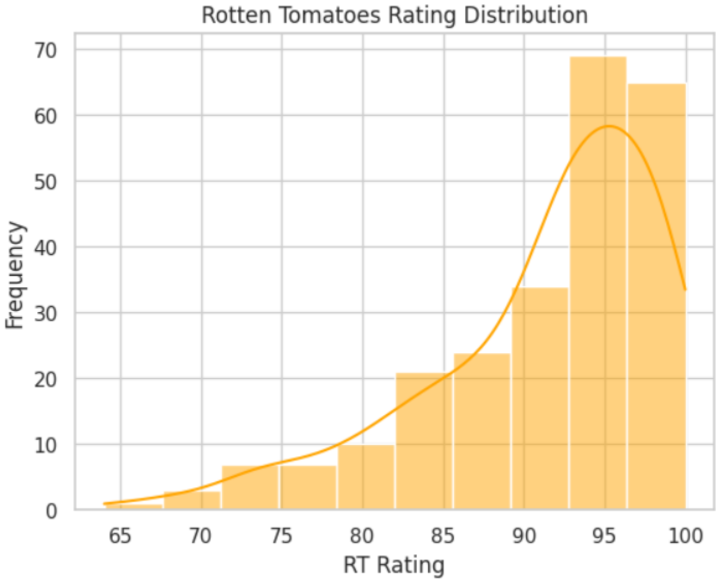
- The dataset was split into training (80%) and testing (20%) sets using `train_test_split`, with a fixed random seed (`random_state=42`) to ensure reproducibility. A `Random Forest Regressor` was selected for model training due to its robustness, ability to handle non-linear relationships, and support for feature importance analysis.
- The model was trained on the training set and evaluated on the testing set using two metrics: **Root Mean Squared Error (RMSE)** and **R-squared (R^2)**. Additionally, the model's `feature_importances_` attribute was analyzed to identify the most influential predictors of IMDb ratings.
- To further improve model performance, a **GridSearchCV** process was optionally applied to fine-tune hyperparameters such as `n_estimators` and `max_depth`. The best parameters were selected based on cross-validation performance.

3. GRAPHS, CORRELATIONS, AND MODEL INSIGHTS (RESULTS AND DISCUSSION)

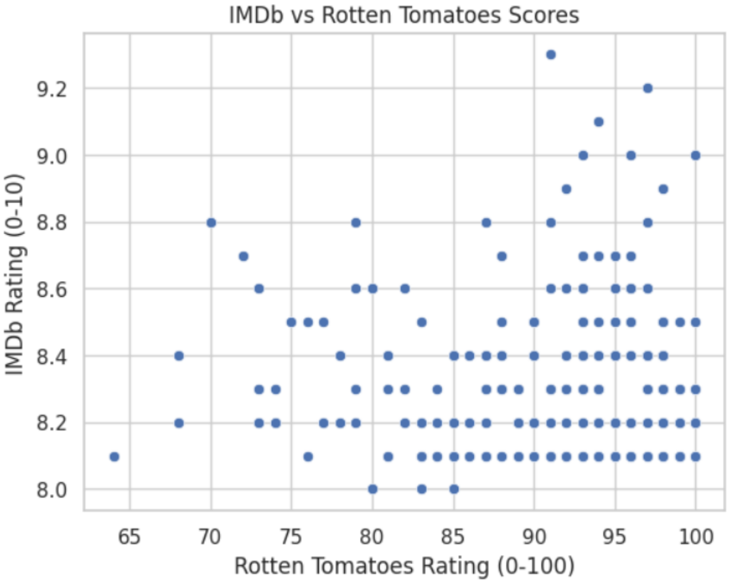
3.1 Exploratory Data Analysis Findings

- IMDb and RT Rating Distributions
  - Histograms were plotted to inspect the distribution of IMDb and Rotten Tomatoes ratings.
  - **Observation**:
  - IMDb Ratings: IMDb ratings tend to cluster mostly between 8.0 and 8.5, with the highest frequency around 8.0-8.2, and the distribution is skewed to the right (positive skew, tail extends to the right). This indicates that the dataset is primarily composed of critically acclaimed or well-received films.
  - Rotten Tomatoes (RT) Ratings: Rotten Tomatoes (RT) ratings (range 65-100) have a much wider spread than IMDb ratings, and the distribution is significantly skewed to the left (negative skew, tail extends to the left) toward higher ratings (especially 90-100).
  - **Understanding**: This suggests that the dataset is biased towards popular or highly rated movies, which may limit the generalizability of the model's predictions to lower rated movies.

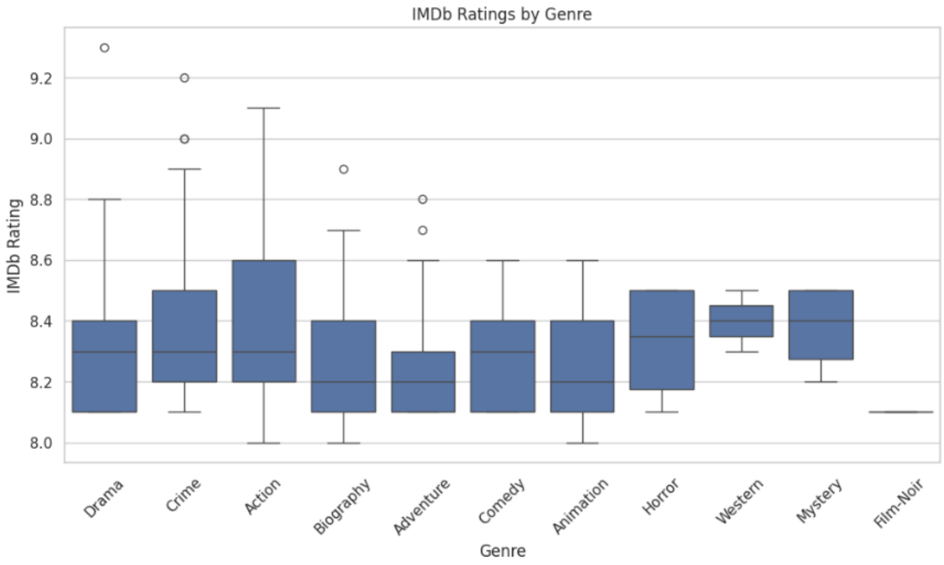




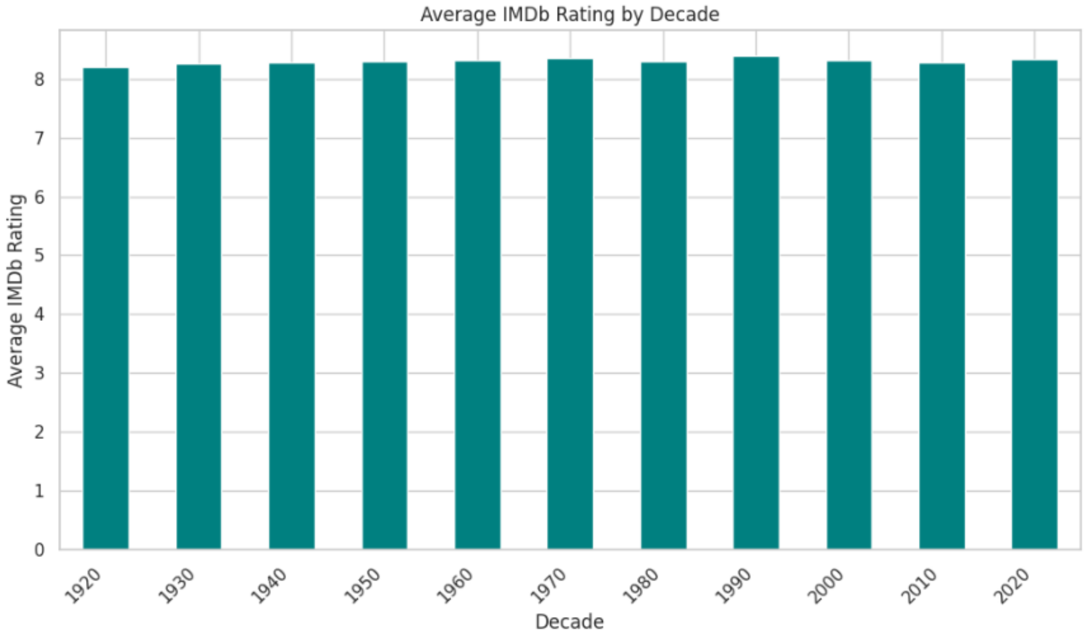
- **IMDb vs. Rotten Tomatoes Scores**
    - A scatter plot was generated to explore the direct relationship between IMDb and RT scores. Additionally, Pearson correlation was calculated.
    - **Understanding**: This correlation implies that higher RT ratings generally align with higher IMDb ratings. However, some outliers suggest occasional discrepancies, which may be due to differing audience and critic sentiments.



- **IMDb Ratings by Genre**
    - Boxplots are available with the MainGenre_Label feature to apply IMDb ratings across genres.
    - **Observation:** Genres like Drama and Crime have higher median ratings than genres like Comedy and Action. The interquartile range (IQR) for these genres (Drama and Crime) is significantly narrower than Action, and similar to the interquartile range for Comedy.
    - **Inference:** This suggests that genre is a meaningful feature in understanding and predicting IMDb ratings.



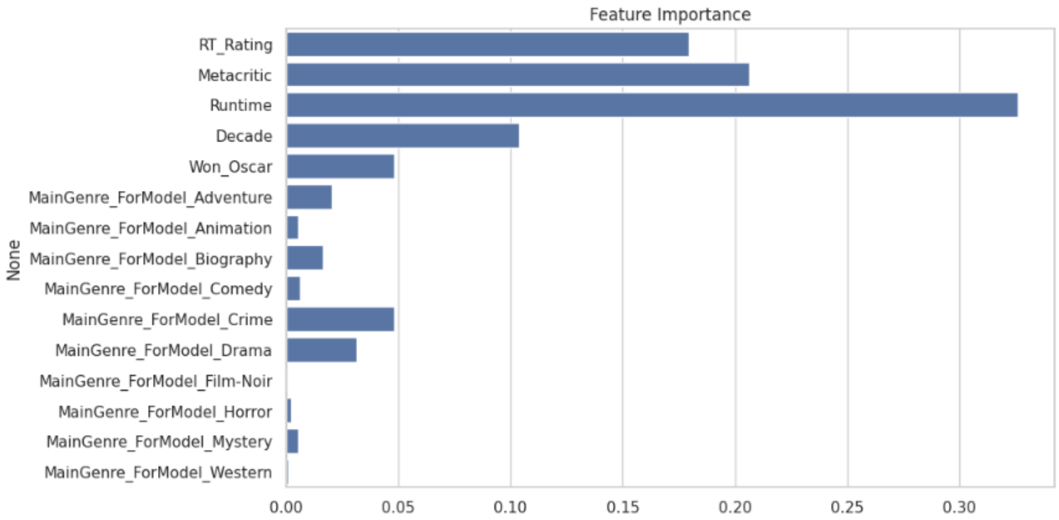
- **Average IMDb Rating by Decade**
  - A bar plot was used to visualize the average IMDb rating by release decade.
  - **Observation**: IMDb scores remain fairly consistent across decades, but there is a **slight increase** in ratings for more recent decades, possibly due to modern production quality or recency bias.
  - **Understanding**: This temporal trend suggests decade could influence public perception and should be retained as a predictive feature.



3.2 Machine Learning Model Performance

- **Initial Model Performance**
  - A Random Forest Regressor was trained using features derived from the dataset. Evaluation metrics were:
  - **RMSE**: ~**0.38**
  - **R² Score**: ~**0.76**
  - **Understanding**:
    The **RMSE of 0.38** indicates that on average, the model's IMDb rating predictions deviate by 0.38 points on a 0–10 scale, which is a relatively low error.
    The **R² score of 0.76** shows that the model explains 76% of the variance in IMDb scores, reflecting good performance and usefulness of the selected features.

- **Feature Importance**
    - A bar chart was created using the feature_importances_ attribute of the model.
    - Observation: The most influential features were:
    - RT_Rating
    - Metacritic
    - Runtime
    - Genre variables (e.g., MainGenre_ForModel_Drama)
    - **Understanding**: This shows that **external critic scores** (RT, Metacritic) are strong indicators of IMDb ratings. Other features such as runtime and genre also contribute meaningfully.



- **Tuned Model Performance (with GridSearchCV)**
    - GridSearchCV was applied to optimize hyperparameters like n_estimators and max_depth.
    - Best Params: {'max_depth': 20, 'n_estimators': 100}
    - Tuned RMSE: 0.223
    - Tuned R²: -0.523
    - **Understanding**: The RMSE of 0.223 suggests that the model's predictions deviate from the actual IMDb rating by approximately 0.22 points on average. However, the negative R² score (–0.523) is a critical indicator that the model is performing worse than a naive baseline model that simply predicts the mean of the target variable.This poor R² score strongly suggests that the selected features do not adequately explain the variance in IMDb ratings. Possible explanations may include:
    - The dataset is too small or biased (e.g., top-rated movies only)
    - The model may be overfitting due to lack of diversity in feature values
    - Important predictive features (e.g., number of reviews, budget, user sentiment) are missing

4. LIMITATIONS AND FUTURE WORK

- **Limitations:**
    - **Model Performance:** The most important limitation is the poor predictive performance of the Random Forest model. A negative R^2 score indicates that the available features may be insufficient or that the model could be trained more.
    - **Dataset Size:** The dataset consists of only 250 movies. A larger and more diversified dataset would make the model more powerful and generalizable.
    - **Feature Set:** The available features are considered. For example, additional data such as the movie's budget, release revenue, director/actor popularity, or more detailed genre representations could provide predictive power.
    - **'Top 250' Bias:** The fact that the dataset is '250 popular movies' may lead to a formation bias. This may limit the generalizability of the model to less popular or lower-rated movies.
    - **Definition of 'Main Genre':** The first specified content of the movie is included as 'Main Genre', a generalized approach, many movies have more than one major genre.

- **Future Work:**
    - **Custom Engineering:** Isolation of more advanced feature engineering; for example, using Natural Language Processing (NLP) on summary stories, incorporating external data that includes communication terms (of their genre and period) or reviews between features.
    - **Alternative Models:** Experiment with other regression applications that can capture more complex relationships (e.g. Gradient Boosting Regressors, Support Vector Regression, Neural Networks).
    - **Extended Dataset:** Collect a larger and more diversified dataset that includes different sources and a wide range of movies.
    - **Error Analysis:** More detailed analysis of where the current model fails to perform (e.g. in specific genres, periods, or rating ranges).
    - **Examining the negative R^2:** The cause of the negative R^2 needs to be understood in more detail. This could be due to data leakage (this is a low case but should be considered), multiple splitting issues (although Random Forest does well in this regard), or issues with selected properties not being as strong as RT/Metacritic in predicting IMDb.
    - **Error Analysis:** "Conduct a thorough error analysis to understand where the current model fails most (e.g., for specific genres, decades, or rating ranges)."
    - **Addressing Negative R^2:** "Investigate the cause of the negative R^2 more deeply. This could involve checking for data leakage (unlikely with this setup but good to keep in mind), multicollinearity issues (though

Random Forest handles this relatively well), or simply that the chosen features do not have a strong enough predictive signal for IMDb rating beyond what RT/Metacritic already provide."

5. FINAL WORDS (CONCLUSION)

- **Key Findings:**
    - IMDb and Rotten Tomatoes ratings show a [e.g., strong positive] correlation, but are not identical, indicating some divergence in how movies are assessed on these platforms.
    - Features like Rotten Tomatoes score, Metacritic score, and genre were identified as important by the Random Forest model for predicting IMDb ratings.
    - However, the developed Random Forest Regressor, even after tuning, exhibited poor predictive performance with a negative R^2 score of [Your R^2], suggesting that it was not able to effectively model IMDb ratings based on the selected features in this dataset.
- **Overall Learnings:**
    - This project provided practical experience in the data science pipeline, from data loading and preprocessing to EDA, model building, and evaluation.
    - It highlighted the importance of feature engineering and the challenge of building accurate predictive models, especially with limited datasets or features. The negative R^2 score underscores that not all datasets will yield highly predictive models with standard approaches and that critical evaluation of results is essential.
    - Further investigation and more complex features or models would be necessary to achieve a satisfactory level of prediction for IMDb ratings.