DS3001 Final Project:
Predicting IMDb Ratings of New Films

Jackson Haiz Quinn Connor Teagan Ryan

Abstract

Movies are a major part of global popular culture, and understanding what drives audience ratings is valuable for filmmakers, streaming platforms, actors, and investors alike. This project develops a predictive model for IMDb movie ratings, using published data from IMDb (the Internet Moive Database). Being able to determine rating predictions based on qualities known before the film comes out gives stakeholders of the film a chance to change plans to increase potential audience reception. We examine how attributes such as genre, runtime, release year, and crew members influence audience scores and implement common machine learning approaches (testing and training) to predict ratings. [RESULTS TO COME AT FUTURE STAGE OF PROJECT]

1. Data Description

IMDb provides publicly available, non-commercial datasets that contain extensive information about films, TV shows, and other media. These datasets include metadata such as release date, runtime, genres, cast and crew, and audience rating. For this project, the key variable of interest is average IMDb rating, which represents the weighted average of user-submitted scores. The object is to build a predictive model that estimates a movie's rating at the time of release, using information available beforehand such as genre, runtime, release year, and the historical success of it's directors and writers. IMDb updates their data frequently, our dataset reflects the version available as of September 26, 2025.

1.1. Variables of Interest

A description of the variables from IMDb.

Identifier / Helper Variables

- tconst (string) - alphanumeric unique identifier of the title
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- numVotes - number of votes the title has received

Target Variable (y)

- averageRating – weighted average of all the individual user ratings

Predictor Variables (x)

- startYear (YYYY) – represents the release year of a title.
- runtimeMinutes – primary runtime of the title, in minutes
- genres (string array) – includes up to three genres associated with the title
- directors (array of nconsts) - director(s) of the given title
- writers (array of nconsts) – writer(s) of the given title

1.2. Cleaning the Data

To prepare the data, we first combined the dataset title.basics.tsv.gz, title.crew.tsv.gz, and title.rating.tsv.gz using tconst as a primary key. Next, we dropped the unneeded columns in title.basics, specifically originalTitle, isAdult, and endYear. From there, we filtered the data to only keep rows where titleType equals “movie,” to exclude TV shows and other types of media. All \N values were converted to NaN for consistency in handling missing data. To increase reliability, we filtered out movies with fewer than 500 votes and then focused our data to only include movies released after the year 2000. This ensures that reviews are more accurate on average and focuses the scope of our model to films from the ”technological era”. After filtering, we are left with 43,816 rows. Finally, we split the data into training and test sets (roughly 80% and 20%, respectively) with the training set including movies released between 2000 and 2020, and the test set including movies from 2020 to 2025.

1.3. Variable Engineering

In addition to the base variables, we created new features to capture historical performance and experience. For directors, we computed the average rating of their past films prior to the release year of the observed movie. For writers, we measured experience by counting the number of movies they had previously worked on. Lastly, because movies can be associated with up to three genres, we created dummy variables for each genre category to allow films to have multi-genre representation.

2. Method Overview

The following outlines methods used to tackle our ratings prediction probelm.

2.1. Method Overview

Our method overview follows the suit of our goals, which is to determine a predictive algorithm that predicts a movie's rating before its time of release, using such factors as genre, runtime, release year, and the historical success of its directors and writers. To determine a correlation of variables, we will first fit our model to enact linear (ridge) regression and random forest. Model performance will be assessed on metrics like RMSE, MAE, and R^2. The data will be split into training and testing subsets to determine each model's validity and guide decisions about what predictors should be included.

2.2. Models Used

Two models were chosen to balance the interpretability of results and predictive complexity.

Ridge Regression (with L2 regularization): Ridge regression serves as the baseline model for our prediction. It is a linear model, but it reduces overfitting and better handles multicollinearity than an ordinary linear regression. This is important as we are aiming to generalize our predictions to new movies, and have some columns that might be related (like "numVotes" and "genre"). Importantly, its simplicity provides a clear interpretation of how each predictor affects the rating.

Random Forest Regression: Random forest regression is a tree-based ensemble method that constructs multiple decision trees and averages their predictions. It can model non-linear relationships and feature interactions, which makes it a tool for capturing complex patterns that might have been missed in our linear model. Feature importance scores will also provide insights to the relative influence of predictors.

2.3. Model Training

The training process will begin with splitting the dataset into training and testing subsets, using an 80/20 split. To simulate real-world conditions, the split will be performed chronologically, with the training set composed of the earliest 80% of titles and the testing set consisting of the remaining most recent 20%. This approach prevents information leakage from future data and more accurately reflects a real prediction scenario, such as ensuring that a sequel does not appear in the training set when its predecessor is included in the testing set. Further, the model parameters will be tuned using 5-fold cross-validation on the training data to identify the predictors that provide the best performance while reducing the risk of overfitting. For ridge regression, this process will involve selecting the optimal alpha level to control the model penalization. For the random forest model, adjustments will focus on key parameters such as the number of trees, the depth of the trees, and how many features are included at each split. Once the best parameters are identified, the final model will be retrained on the full training set and then evaluated on the test data to measure how well it can predict the rating of new films.

2.4. Model Validation and Implementation

To ensure robustness and generalizability, we will perform k-fold cross-validation on the training data. Additionally, evaluation metrics of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R^2 will be used to assess model performance. We compare Ridge Regression and Random Forest Regression models based on these metrics to determine which best predicts a movie's rating prior to release. A confusion matrix will also be generated for ratings binned into categories (like 0-1, 1-2, . . . , 4-5) to further assess classification-style performance.