Predicting Movie Ratings

Movie Money Matters, jsfarrel, bkaravan, ystepane

Goal

A movie's IMDb rating is a quantitative measure of success, but will not always correlate to other benchmarks such as box-office gross. The use case for this project is to determine underlying factors that influence a movie's rating. Underlying factors include statistics such as the writer's name, runtime, budget, gross, genre, release date. Thus, we are most interested in the following: given a movie's statistics, predict the movie's average IMDb rating.

Data

Our dataset consists of various movie attributes and their average IMDb ratings. We combined existing IMDb datasets with web scraped box-office statistics from The Numbers. Our dataset consisted of 10,715 entries, reduced to 1,311 after cleaning. Cleaning included removing duplicate movie entries and entries with missing data points. Preprocessing included formatting dates and monetary values so they are uniform across the dataset and easy to analyze later on. Our final dataset had 14 attributes, including movie name, budget, domestic gross, worldwide gross, date (day, month, year), writers, directors, genres, runtime, number of votes, and average IMBd rating.

Model+Evaluation Setup

We trained Elastic Net and Random Forest models to predict the rating of a movie based on the attributes we have in our dataset. We chose to represent movie ratings as numerical continuous values, rather than categorical, therefore making this a regression problem. We believed a categorical representation may be an oversimplification of the problem. The target variable was IMDb average rating, and attributes included everything but the movie's title. During preprocessing, columns were transformed using a one-hot encoder for categorical values, ordinal encoder for ordinal values, and standard scaler for numerical values. We used a K-fold cross validation pipeline consisting of 5 splits to tune hyperparameters. The models were evaluated using mean squared error to penalize large errors and prevent overfitting of the data. We performed an 80%-20% train-test split of the dataset.

Results and Analysis

Claim #1: Both regression models outperform the baseline by a significant margin.
Support for Claim #1: Table 1 shows the MSE of Random Forest and Elastic Net models and the baseline on the test dataset. Both models MSE < 0.850, while the baseline > 1.100.

Model	**Mean Square Error (MSE)**
Random Forest	0.596
Elastic Net	0.808
Baseline	1.145

Table 1: MSE of different models on testing dataset

Claim #2: The relationship between our movie dataset attributes and movie rating is non-linear.

Support for Claim #2: Random Forest is a non-linear model, while Elastic Net is a linear regressor. The relationship between the attributes in our dataset and movie rating is non-linear, because as Table 1 (above) shows, our movie rating predictions are best with non-linear Random Forest compared to Elastic Net and the baseline.

Claim #3: Number of votes and runtime are the two most important factors for a movie's rating.
Support for Claim #3: After determining the best model from Table 1, Random Forest, we ranked features by permutation importance and global feature importance in Figures 1 and 2, respectively. The features that stand out as being most important globally and due to permutations include runtime in minutes and number of votes.

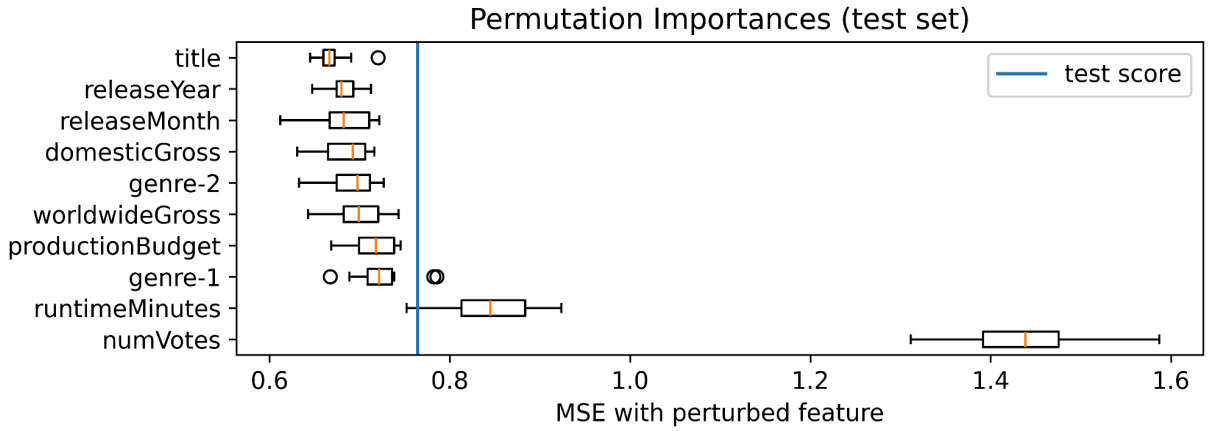

Figure 1: Permutation importance for Random Forest. Higher MSE means greater significance.

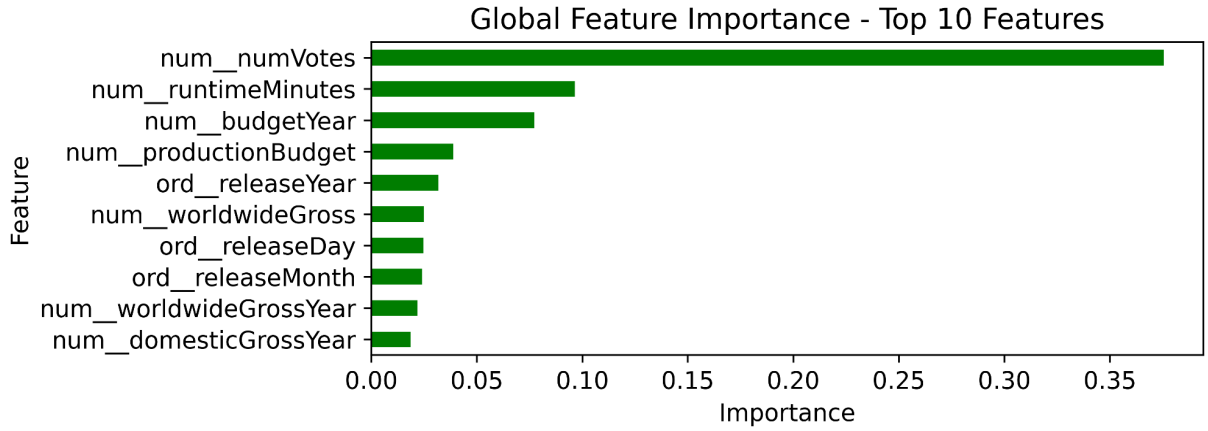

Figure 2: Random Forest global feature importance.