Movie Math

mchang35, ebussman, ebelt1

Goal

The film and entertainment industry has only been growing, especially with the increase in streaming activity over the past year. By better understanding what makes a movie popular or successful, directors can make movies that better fit the desires of the masses. Therefore, we are interested in predicting which movies are likely to be highly rated by the population and to bring in a high profit based on characteristics such as associated keywords, genre, casting, directors, production companies, and production country.

Data

The data was collected from Kaggle, a subsidiary of Google that allows data scientists to collect and share datasets. Two different sets were used: 1) "The Movies Dataset," which consisted of metadata and keyword associations for 45,000 movies released during or before July 2017, and 2) "IMDB Movies Extensive Dataset," which consisted of metadata and IMDB user rating information for over 85,000 movies released between 1874 and January 2020. The datasets were joined based on IMDB ID, a unique identifier for each movie. We were left with a movies metadata table, including over 32,000 movies, as well as a keywords table, including over 150,000 movie/keyword pairs.

Model and Evaluation Setup

We used two methods for analyzing the relationship between a movie's characteristics and its average user rating and profit: Naive Bayesian Classifiers and multiple regression. We used two separate Naive Bayesian Classifiers to predict whether a movie was going to be above or below the 70th percentile of profit and average user rating based on its keywords. Keywords included both movie features (such as "female director" or "post-credit scene") and movie themes (such as "friendship," "war," or "Paris"). We randomly split the data 80/20 into training and testing data, respectively, and measured the success of these algorithms with their prediction accuracy and false positive and false negative rates. We used two multiple regressions to predict the average rating and profit of a movie. The independent variables used were runtime, number of popular actors, whether the movie was directed by, written by, and produced by popular directors, writers, and production companies, country of production, original language, season of release, whether the movie was in a series or not, and genres. Each model had 33 independent variables in total (including indicator variables). We measured the success of this algorithm by comparing the r-squared and RMSE values of the training data (80% of the data) and the testing data (20% of the data).

Results and Analysis

Claim #1: Our Naive Bayesian Classifiers, using the 70th percentile thresholds, outperformed the baseline accuracy of 50% by a sizable margin.

Support for Claim #1: The table below displays the accuracy rate of the Naive Bayesian Classifier in determining whether movies are above or below the 70th percentile average rating and profit thresholds. Both accuracy rates are greater than the 50% baseline by approximately 25 percentage points.

	Mean Vote	**Profit**
Accuracy	0.7646818088807039	0.7452547452547452

Claim #2: Our multiple regressions are able to make relatively accurate predictions for movies beyond our training dataset about movies' average rating and profit based on key features.
Support for Claim #2: The table below displays the r-squared value and RMSE value for the training data and testing data for both models. For both models, the training and testing r-squared values were about the same, with differences less than 0.0322. The RMSE values were either similar or less for the testing data compared to the training data for both the mean vote and profit models. Therefore, our models can be generalized to predict the success of other movies beyond the training data.

	Mean Vote		**Profit**	
	Training	Testing	Training	Testing
R^2	0.291	0.3155261979	0.313	0.2809082414
RMSE	0.8042589543	0.8184920003	120694839.7	102815650.8

Claim #3: Possessing various characteristics, such as certain genres, being directed by a popular director, country of production, and more can increase the expected profit and mean rating of a movie.
Support for Claim #3: Statistically significant positive relationships were found between mean vote and certain genres (such as animation, film-noir, drama, and biography), directed by popular directors, and written by popular writers. Statistically significant positive relationships were found between profits and certain genres (such as animation, sci-fi, adventure, and fantasy), number of popular actors, produced by popular production companies, country of production, and whether the movie was a series or not.
The tables below display the coefficients of the independent variables in the models predicting the profit and mean rating of movies. All coefficients are positive, which means that if a movie has that characteristic, the profit and mean vote of movies are expected to increase. These relationships are statistically significant at the 99% confidence level, since all p-values are less than 0.01. A p-value less than 0.01 means that if there was no relationship between the independent and dependent variables, data as extreme or more extreme than the training data observed would occur less than 0.01 of time (which is super rare). Therefore, these p-values provide evidence that we can reject the notion that there is no relationship between these variables and the dependent variables of mean vote and profit.

Profit Relationships

	Coefficient	**P-Value**
Animation (Genre)	67.8 million	<0.001
Sci-Fi (Genre)	25.8 million	0.001
Adventure (Genre)	51 million	<0.001
Fantasy (Genre)	38 million	<0.001
Number of Popular Actors	1.8 million	0.004
Produced by Popular Production Company	21.6 million	<0.001
Produced in the US	54.8 million	<0.001
Series	94.5 million	<0.001

Mean Vote Relationships

	Coefficient	**P-Value**
Animation (Genre)	0.5148	<0.001
Film Noir (Genre)	0.9982	<0.001
Drama (Genre)	0.2855	<0.001
Biography (Genre)	0.1628	0.004
Directed by Popular Director	0.2275	<0.001
Written by Popular Writer	0.3327	<0.001

**Note: Other statistically significant relationships were found but not discussed