Movie Math

An Analysis on What Makes a Film Successful

ebelt1, mchang35, ebussman

Background

In a world of blockbuster bests, such as "Avengers: End Game," and film flops, such as "Cats," we aim to find what makes a film successful.

Prediction Task

We are interested in predicting which movies are likely to be highly rated by the population and generate high profit based on characteristics such as associated keywords, genre, actors, directors, production companies, writers, runtime, release date, and languages.

Data Collection

- We joined two datasets found on Kaggle, a subsidiary of Google used for sharing datasets.
  - "The Movies Dataset" was originally extracted from TMDB and GroupLens, movie rating and recommendation websites
  - "IMDB Movies Extensive Dataset" was scraped from IMDB, a movie review site
- The final joined dataset included 33,748 movies with 39 total attributes, as well as 155,991 movie/keyword association pairs.

Methodology

Naive Bayes Tests Based On Keywords

- We used Naive Bayes to construct a model to analyze whether a movie would be successful based on its associated keywords.
- Associated keywords included movie features (such as female director or post-credit scene) and movie themes (such as friendship or Paris)
- We used two separate Naive Bayes Classifiers: one based on profit and the other on mean vote
- Movies were labeled as successful if they were above the 70th percentile profit (72 million) or 70th percentile average rating (7.2).
- The data (movie/keyword pairs) was split 80/20 into training and testing data, respectively.

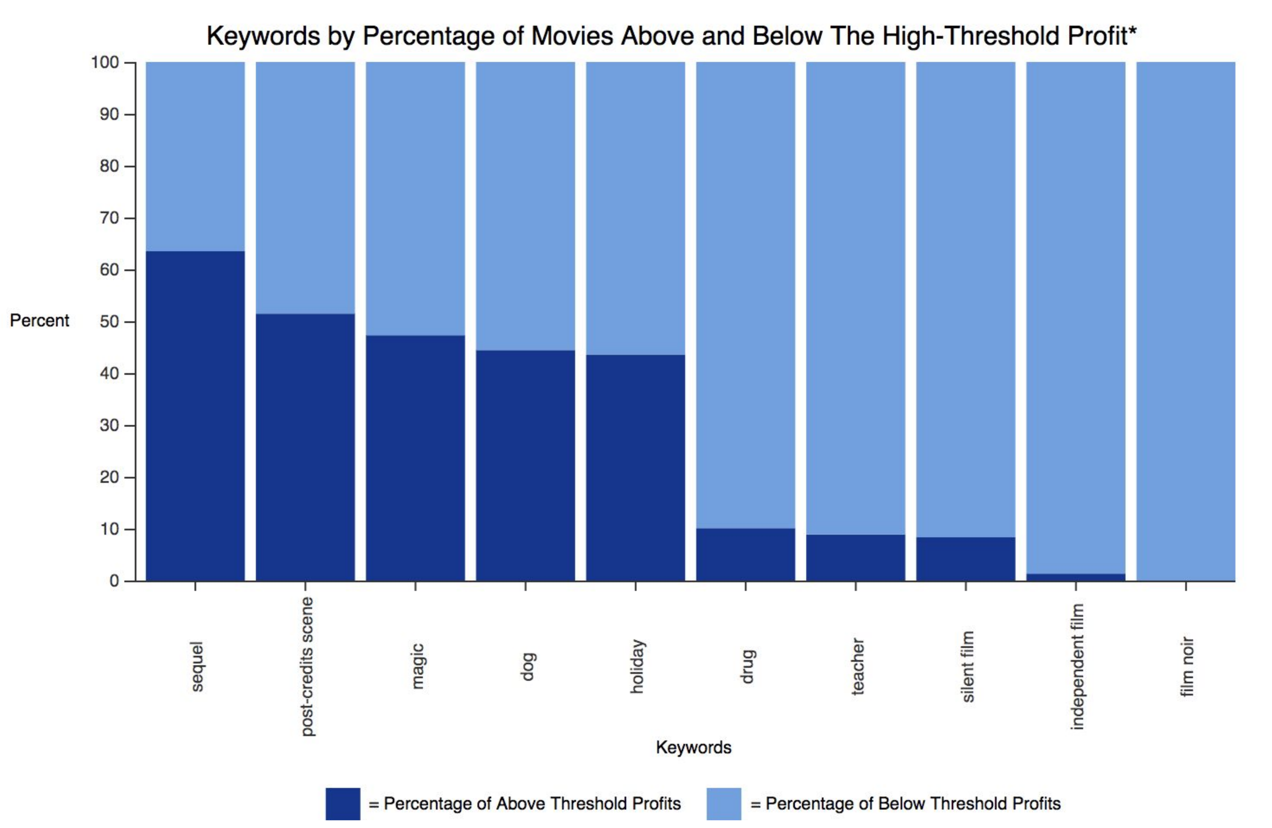
Regression

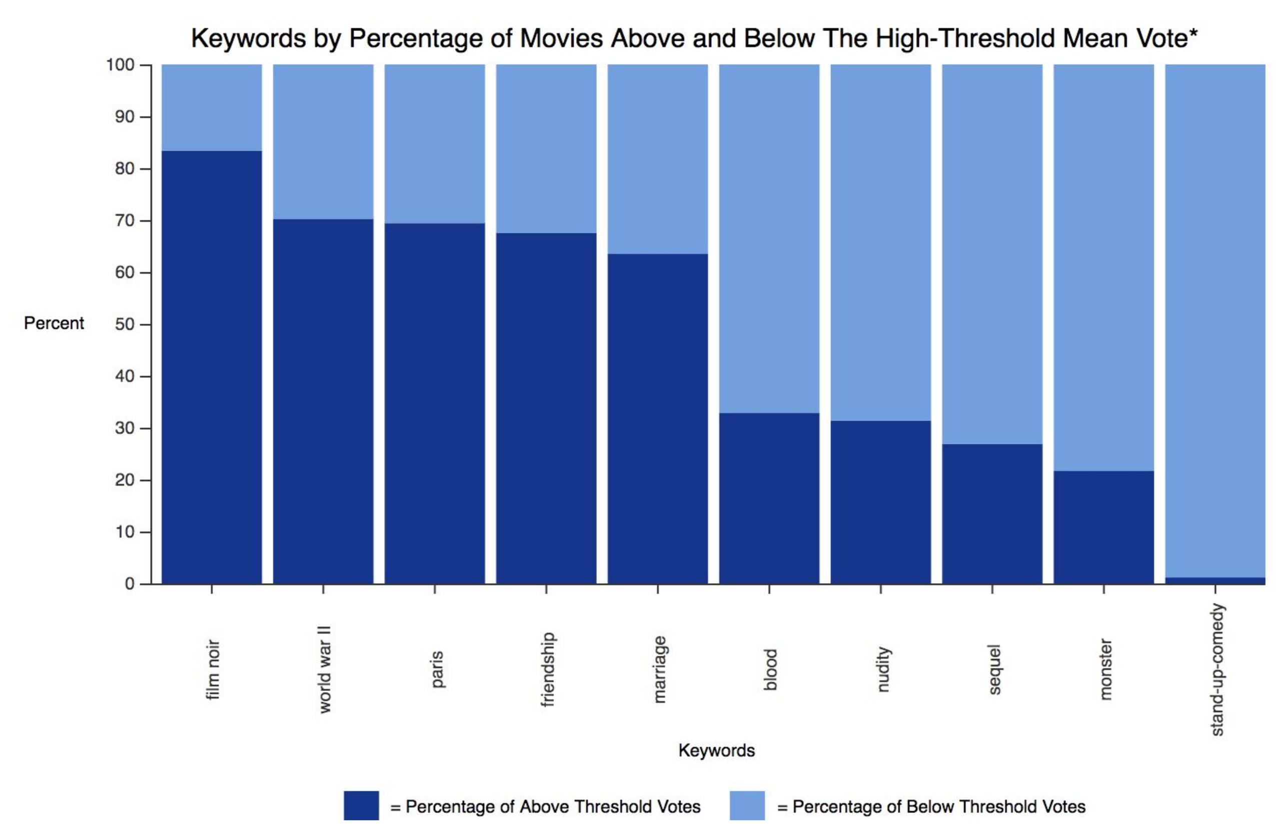
- We used an ordinary least squares multiple regression model to predict mean user rating and movie profit
- The data (movies) was split randomly 80/20 into training and testing data, respectively.
- Independent variables include: runtime, popular actors, popular directors, popular writer, use of popular production company, country of production, original language, release date, whether the movie was a series or not, and genre
- Each model had 33 independent variables in total (including indicator variables)

Naive Bayes Results

	Mean Vote	Profit
Percentile	0.70	0.70
Threshold	7.2	71971200.0
Number of Training Data Points (Number of Movies)	19528	3921
Number of Testing Data Points (Number of Movies)	4887	1001
Probability that Y = 1	0.23120647275706677	0.24483550114766642
Accuracy	0.7646818088807039	0.7452547452547452
False positive rate (predict Y=1 when Y=0)	0.10644855381697486	0.13914027149321267
False negative rate (predict Y=0 when Y=1)	0.38537658053875756	0.3548387096774194

- Of the top 30 most-frequently used keywords (in at least 250 movies), the two visualizations below display the five keywords with the highest percentage of films above the threshold and the five keywords with the lowest percentage of films above the threshold.
- "Sequels" had the highest percentage (63.5%) of movies above the high profit threshold. "film noir" and "stand-up comedy" had the lowest percentages (0%) of movies above the high profit threshold.
- Corresponding to this finding, "stand-up comedy" also had the lowest percentage (1.15%) of films above the high mean-vote threshold. Interestingly, "film noir" had the highest percentage (83.4%) of movies above the high mean vote threshold, which suggests that good reviews do not always lead to high profits.



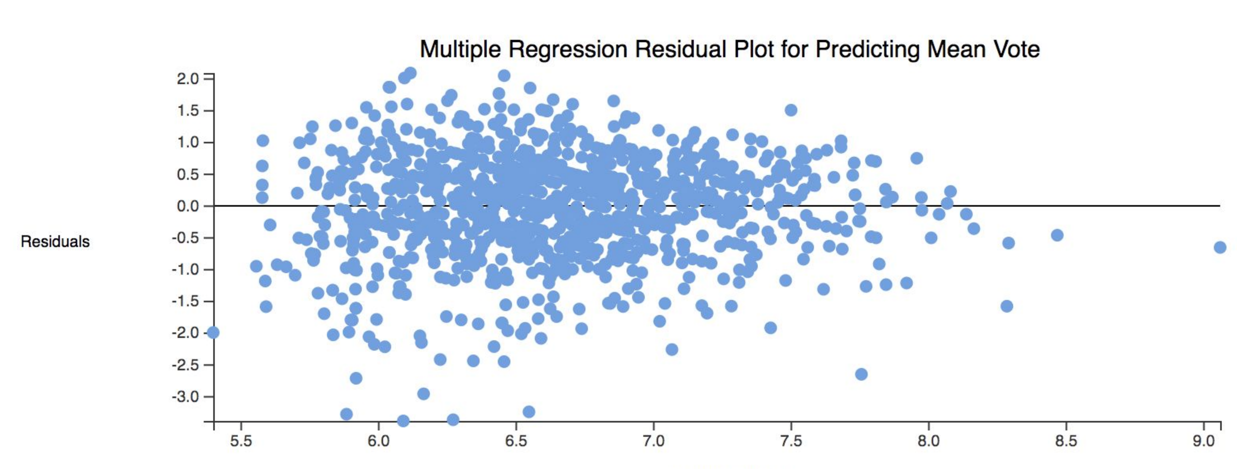


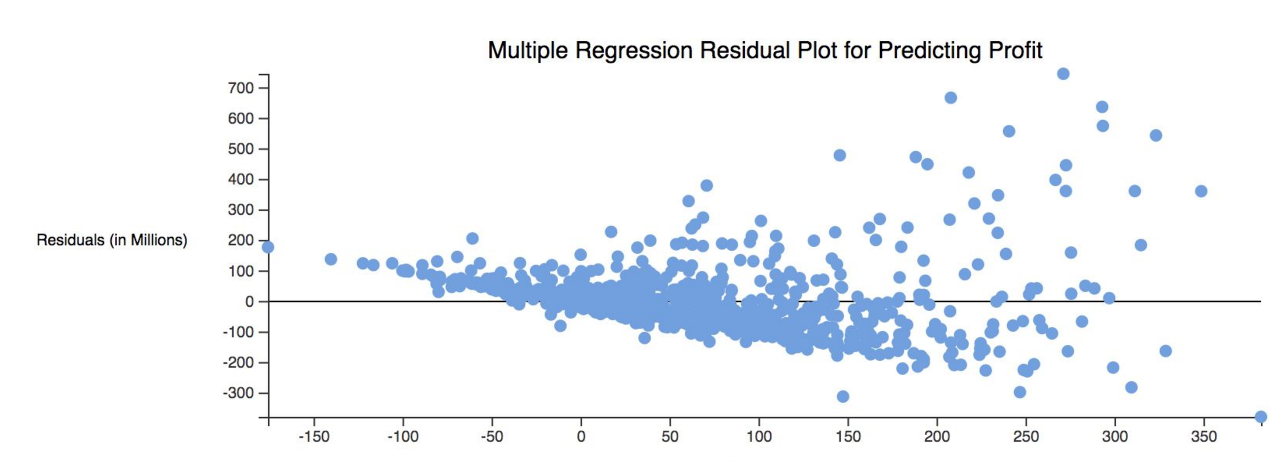
Note: the tests used thresholds at the 70th percentile, the graphs used thresholds at the mean for better visualization.

Regression Results

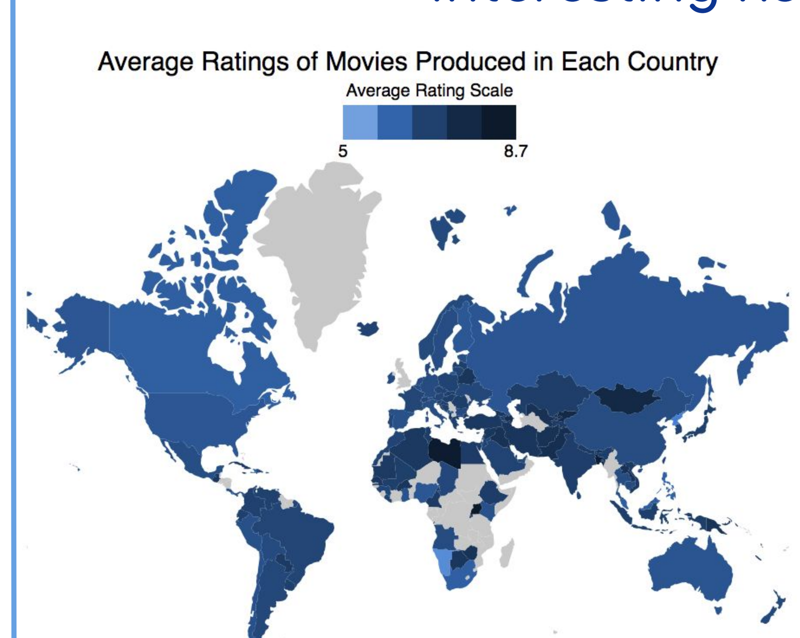
- Plots showing the residuals vs. the predicted values of the testing data for both the mean vote and profit models are shown below
- The residual plot of the mean vote model has more evenly-dispersed residuals in terms of positive and negative values than the residual plot of the mean profit model. This distribution means that the linear regression model is a better fit for predicting mean vote than for predicting profits
- In the table, the R^2 values for the training and testing data are about the same (in both models), which indicates that the models can be used just as well to predict the success of movies beyond the training data

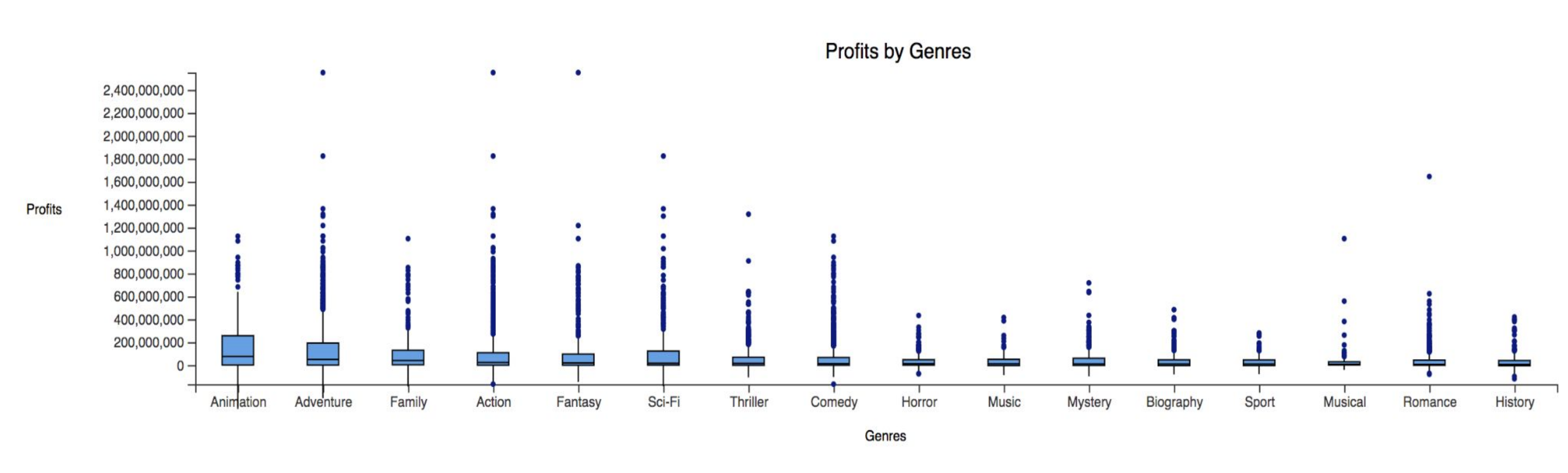
	Mean Vote	Profit
Train R^2	0.291	0.313
Test R^2	0.31552619793130476	0.28090824137613224
Train RMSE	0.80425895426	120694839.68
Test RMSE	0.81849200027	102815650.766





Interesting Relations Between Success and Independent Variables





Evaluations and Takeaways

Naive Bayes Test on Keywords

- Our Naive Bayesian Classifiers had accuracy rates of 74%-76% (above the baseline performance of 50%), with false positive rates of 10%-13% and false negative rates of 35 - 38%.
- An example of the model's correct label of "successful," in terms of profit, is "Star Wars: Episode VII - The Force Awakens" (which had profits more than 10 times the threshold). Its keywords were "iMax," "jedi," "space opera," "3D," "android," and "spaceship": 87%, 86%, 76%, 66%, 46%, and 46% of movies with these keywords, respectively, had profits above the threshold.
- An example of the model's correct label of "unsuccessful," in terms of profit, is "King Arthur: Legend of the Sword" (which had negative profits (i.e. it lost money)). Its keywords were "3D," "period drama," "King Arthur," and "sword": 34%, 54%, 60%, and 75% of movies with these keywords, respectively, had profits below the 70th percentile threshold.

Regression

- Our multiple regressions had testing R^2 values between 0.28 and 0.316
- Statistically significant positive relationships were found between mean vote and certain genres (such as animation, film-noir, drama, and biography), inclusion of popular directors, and inclusion of popular writers.
- Statistically significant positive relationships were found between profits and certain genres (such as animation, sci-fi, adventure, and fantasy), inclusion of popular actors, use of popular production companies, country of production, and if the movie was a series or not

Relevant Limitations

- The Naive Bayesian Classifier assumes that the keywords of each movie are independent from one another. This assumption, however, is likely not to hold for our data, since at least some keywords are likely to be related to each other.
- The popular actors, writers, production companies, and directors attributes are based on the total number of movies, within our dataset specifically, of which these figures were part. For a given movie, it does not account for how many movies the figure was previously in at the time the movie was made (i.e. if the figures were not popular yet). Additionally, these attributes do not account for if these figures are popular from roles outside of movies (such as television)
- The range of mean user ratings for movies was somewhat narrow (approx. 5/10 -- 8/10) and profit data (revenue, budget) was relatively sparse.