PS11: Rough Draft

Audrey Hopewell

April 14, 2020

1 Introduction

This project is inspired by FiveThirtyEight's Hollywood Taxonomy series, in which an actor's movies are classified into 3-5 categories by plotting their box office gross against their Rotten Tomatoes rating. This results in categories like "The Most Cowbells" for Will Ferrell (high grossing and highly-rated, such as The LEGO Movie) [10] and "Hope Sinks" for Sandra Bullock (low grossing and low-rated, such as Miss Congeniality 2)[11].

These categories are humorous and provide retrospective insight, but is there a way to make this concept more useful? For example, is there something about these movies themselves (i.e. their characteristics before they are filmed, edited, and released) that predicts which category they'll fit into? If so, knowing these characteristics could help actors make more informed decisions about whether a movie will become a beloved but unprofitable cult classic or a box office smash with terrible reviews.

2 Literature Review

- Actors are likely interested in both revenue and reviews because Hollywood is a highly reputation-based industry. A history of appearing in well-received and profitable films positively affects an actor's ability to negotiate for future desired projects [5]
- Hollywood has been characterized as a project-based industry because each film is made by a novel group of cast and crew. The film itself is a project firm because it comes together, creates an output, and then dissolves [8][12]. The nature of the industry means that actors constantly negotiate new employment agreements based on the success of past projects.
- There is a lot of machine learning literature attempting to predict movie revenue or profitability based on pre-release data (e.g. cast and director, release season, or text analysis of the script), post-release data (e.g. social media sentiment) and econometric work analyzing the relationship between movie characteristics and revenue [13][2][3][6][7][14][15][9][1]. This work typically includes variables related to the starring actors, such as

genre expertise[13] or historical average revenue[6], with the aim of helping investors make more informed decisions about funding the movie after the main cast is already involved.

- Typical measures of movie success include profit[13], revenue[1][9][14][15], or individual and collective awards[4]
- Given this, how should actors decide whether to sign on to a movie?

3 Data

- I use a movie metadata set from Kaggle, which includes information compiled from IMDb's various publicly available data sets. This includes the following variables:
  - whether the movie is "adult"
  - which "collection" or series it belongs to, if any (e.g. Toy Story)
  - its budget
  - its genre(s)
  - its language
  - its original title
  - production company
  - production country
  - release date
  - revenue
  - runtime
  - which languages are spoken in the film
  - status (e.g. released, in production)
  - its tagline
  - its title
  - average IMDb rating
  - how many ratings it has received on IMDb
  - an overview (summary) of the movie
- For simplicity's sake, I will use IMDb rating as a substitute for FiveThirtyEight's use of Rotten Tomatoes. Rotten Tomatoes is the more popular ratings site, but IMDb's data is more easily downloadable and already part of the data set.

- I clean the data by omitting movies that are considered adult (as I only want mainstream movies that would be released in theaters),movies that are not in the English language, those that were not produced in the United States, and movies that went straight to video. I limit the data to only movies that have been released (otherwise box office and rating data would not be available). I also eliminate movies released before 1990 and movies for which revenue is listed as 0.
- After cleaning, 4128 movies remain for analysis.

4 Empirical Methods

- Dependent variable is which category the movie falls into. The assumption is that for an actor's reputation, the general reputation of a movie (as represented by its quadrant on the graph of revenue vs. IMDb rating) is more important than the exact revenue or popularity numbers of the movie. As the literature demonstrates, movie success cannot be explained fully by the actors themselves[13], so general movie reputation would seem to be more important.
- Categories are formed by plotting all movies (after data cleaning and filtering) by their revenue and IMDb rating, then splitting them into four categories based on the mean value of each variable. Thus, a movie is "high revenue" if it takes in more revenue than average and "highly rated" if it is more highly rated than average.
- Independent variables are as follows:
  - dummy variable for whether the movie belongs to a collection (i.e. is part of a series)
  - movie budget
  - dummy variables for each genre
  - dummy variable for each production company
  - dummy variable for each additional production country other than the U.S.
  - dummy variable for each release season (this is an imperfect variable because release may be rushed or delayed and so actual release season might be different from planned release season. However, it is included since most movies should be released around the time they were originally planned to be)
  - dummy variable for each additional spoken language other than English
  - I'm considering doing sentiment analysis on the "overview" of each movie and using the prominence of each sentiment as a variable. This could approximate the tone of the movie beyond genre.

- Then, I will try to determine which machine learning algorithm will most accurately predict the outcome - movie category - given the extremely basic variables provided. I'll test:
  - Trees
  - Logistic regression
  - Naive Bayes
  - kNN
  - SVM
- Is there a way to make the prediction more accurate by including information beyond the movie's metadata? Reputation is important for actors looking for jobs, but a director's reputation is also key in attracting actors to a project since they're more likely to be signed on before the cast[6]. Using a cast and crew data set, I extract the director for each movie in my data set. Then, I construct a very basic measure of "director reputation" similar to the movie reputation measure: I assign them the category into which the plurality of their movies fall. This director reputation is then included as an independent variable.
- We can expand the above idea by also creating a writer reputation variable, in which each writer is assigned the category into which the plurality of their movies fall.
- We can also analyze the movies based on profitability (rather than revenue) and rating. I construct a new variable, profitability, equal to the different between revenue and budget, and then create profit categories. "High" profit is greater-than-mean profit. Then, I will re-run the algorithms to see:
  - which does the best job at predicting the new category
  - whether profit-based categories are easier or harder to predict than revenue

5 Research Findings

6 Conclusion

Potential weaknesses and limitations:

- IMDb ratings might be biased (e.g. people who are film buffs are more likely to use IMDb)
- Director reputation obviously depends on more than the reputation of their past movies (e.g. are they difficult to work with). This type of information spreads through the concentrated social network of the Hollywood film industry[5] but is impossible to measure with the data we have.

References

[1] K. R. Apala, M. Jose, S. Motnam, C. C. Chan, K. J. Liszka, and F. de Gregorio. Prediction of movies box office performance using social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, page 1209–1214, 2013.

[2] M. Baimbridge. Movie admissions and rental income: the case of james bond. *Applied Economics Letters*, 4:57–61, 1997.

[3] P. Boccardelli, F. Brunetta, and F. Vicentini. What is critical to success in the movie industry? a study on key success factors in the italian motion picture industry. *Dynamics of Institutions and Markets in Europe*, 46, 2008.

[4] G. Cattani and S. Ferriani. A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the hollywood film industry. *Organization Science*, 19(6):824–844, 2008.

[5] J. J. Ebbers and N. M. Wijnberg. Disentangling the effects of reputation and network position on the evolution of alliance networks. *Strategic Organization*, 8(3):255–275, 2010.

[6] A. Elberse. The power of stars : Do star actors drive the success of movies? *AMA Journal of Marketing*, 71(October):102–120, 2007.

[7] J. Eliashberg, S. Hui, and Z. Zhang. Assessing box office performance using movie scripts: A kernel-based approach. *Knowledge and Data Engineering*, 26(11):2639–2648, 2014.

[8] S. Ferriani, G. Cattani, and C. Baden-Fuller. The relational antecedents of project-entrepreneurship: Individual connectedness, team composition and project performance. In *Academy of Management Proceedings*, volume 1, pages 1–6, 2008.

[9] S. Gopinath, P. K. Chintagunta, and S. Venkataraman. Blogs, advertising and local market movie box office performance. *Management Science*, 59(December):2635–2654, 2013.

[10] Walt Hickey. The four types of will ferrell movies, 2015.

[11] Walt Hickey. The three types of sandra bullock movies, 2015.

[12] C. Jones and K. Walsh. Boundaryless careers in the us film industry: Understanding labor market dynamics of network organizations. *The German Journal of Industrial Relations*, 4(1):58–73, 1997.

[13] M. T. Lash and K. Zhao. Early predictions of movie success: the who, what, and when of profitability. *Journal of Management Information Systems*, 33(3):874–903, 2016.

[14] R. Parimi and D. Caragea. Pre-release box-office success prediction for motion pictures. In *Machine Learning and Data Mining in Pattern Recognition*, pages 571–585, 2013.

[15] J. S. Simonoff and I. R. Sparrow. Predicting movie grosses : Winners and losers , blockbusters and sleepers. *CHANCE*, 13(3):15–24, 2000.