Predicting Movie Outcomes by Pre-production Characteristics

Audrey Hopewell*

May 4, 2020

Abstract

This paper uses six machine learning algorithms to predict movie success using pre-production characteristics. Movie success is defined as the relationship between revenue and audience reception, which is operationalized as user-generated 1-10 ratings. Movies are classified using two different methods: an "intuitive" manual classification scheme using the mean of each outcome variable to split the data into quadrants, and k-means clustering. Then, the six machine learning algorithms perform multi-class classification for both types of classification methods. I find that the classes generated by k-means clustering facilitate more accurate predictions than "intuitive" classes and that the support vector machine is the most successful algorithm for this method.

*Department of Economics, University of Oklahoma. E-mail address: audrey.hopewell@ou.edu

1 Introduction

The great writer and director Sidney Lumet writes in "Making Movies" that "I don't know what makes a hit. I don't think anyone does" (Lumet, 1995, pp. 198). This sentiment might be familiar to anyone who spends $10 on a ticket to see a movie starring their favorite actor, directed by a lauded filmmaker, for which the trailer looked fascinating, and ends up wanting their money and time back. Conversely, some movies that flop in theaters are revered as masterpieces of the craft. If even experts like Lumet don't know what makes a good movie, perhaps there is a chance for data analysis to shed some light on the relationship between the movie itself and its performance.

This project is inspired by FiveThirtyEight's Hollywood Taxonomy series, in which an actor's movies are classified into three to five categories by plotting their box office gross revenue 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) and "Hope Sinks" for Sandra Bullock (low grossing and low-rated, such as Miss Congeniality 2) (Hickey, 2015b,a).

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? Does a certain genre tend to produce beloved but unprofitable cult classics or box office smashes with terrible reviews, for example? If so, knowing these characteristics could help moviemaking professionals (actors, editors, production designers, etc.) or investors predict a movie's likely performance (in revenue and audience reception)

and thus make more informed employment or funding decisions.

2 Literature Review

Previous research in the machine learning, management sciences, and economics fields has examined the relationship between movie profitability and movie characteristics. Such research has addressed both movie metadata (such as genre, budget, etc.) and more sophisticated metrics (such as measures of the social networks and expertise of a movie's cast and crew). Key factors in movie success have been defined as audience-based (such as movie reviews or social media sentiment), release-based (such as the number of theaters a movie is released in or the time of year of release), and movie-based (characteristics of the film itself, such as genre and cast information) (Lash and Zhao, 2016). Movie-based characteristics are the only ones available before release, and only a subset of that information is available before a movie is made (e.g. the run time or MPAA rating is not known until the movie has been produced).

In machine learning, the research typically focuses on predicting the success of a movie based on its characteristics. Sharda & Delen, for example, take a classification approach, using pre-release characteristics and a neural network to predict which of nine revenue-based categories into which a movie will fall (Sharda and Delen, 2006). Parimi & Caragea expand this work by including network-based features for revenue classification to account for the dependencies among movie outcomes, e.g. the effect of a star director or actor (Parimi and Caragea, 2013). Lash & Zhao take a similar approach of using pre-release characteristics–including novel network measures like the "genre expertise" of a cast weighted

by each cast member's "star power" (i.e. average profitability) and actor-director collaboration frequency–but focus on predicting movie profitability and return-on-investment (ROI) rather than revenue, with the goal of constructing a tool for investors to make funding decisions (Lash and Zhao, 2016). Other work, such as that by Elberse, finds that major casting announcements–a pre-production factor–are important drivers of both immediate movie valuation and theater revenue (Elberse, 2007).

Some innovations in the literature include the use of novel measures of star power, movie content, or cultural "hype" around a film. Apala et al, for example, mine data from social media and use the number of director and cast Twitter followers, the number of trailer views and comments on YouTube, and a sentiment analysis of those YouTube comments to predict movie revenues (Apala et al., 2013). Eliashberg et al use a kernel approach to analyze movie scripts, and using this and estimated production budget (capturing only features known when a movie is "green-lit"), are able to predict box office revenue more accurately than other measures (Eliashberg et al., 2014).

Movie success is typically measured by revenue (Apala et al., 2013; Parimi and Caragea, 2013; Simonoff and Sparrow, 2000; Gopinath et al., 2013; Sharda and Delen, 2006; Elberse, 2007; Eliashberg et al., 2014). However, some research has used profit or return-on-investment (ROI) as an alternative measure that is meant to be more relevant to investors (Lash and Zhao, 2016). In non-machine learning literature (e.g. management or organizational sciences, who are more focused on the nature of creative processes and success), movie success is measured by individual or collective awards won by the cast and crew (Cattani and Ferriani, 2008). Machine learning literature has not typically incorporated this measure directly as a dependent or independent variable, as "star power" measures usually include

revenue/profit metrics or social media following (Lash and Zhao, 2016; Sharda and Delen, 2006).

There is a lack of literature focusing on audience reception as an output. While some (Apala et al., 2013) use audience reception as a predictor of movie revenue, the literature typically treats audience opinion as an input to financial indicators rather than something that itself can be predicted or is an outcome of interest. Intuitively, movie professionals are likely interested in audience and peer perception just as much as they are in movie revenue or profit, as movie reception speaks more directly to their talent than revenue or profit, which may simply reflect the size of the advertising budget.

Indeed, 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 (Ebbers and Wijnberg, 2010). Additionally, there is a positive feedback loop, as professional crew members with a history of award success tend to work with other frequently-award professionals and produce more creatively-lauded films Cattani and Ferriani (2008). Thus, it is in a movie professional's interest to sign on to films that will be creatively well-received, in addition to being profitable.

Even more basically, a strong reputation of creative success is necessary in forming the kinds of social networks that facilitate sustainable employment in the movie industry. 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 Ferriani et al. (2008); Jones and Walsh (1997). Project firms form largely based on social network connections, which allow information–particularly about skill and reputation–to travel between nodes (individuals in the network) (Jones and

Walsh, 1997). Thus, a history of working on movies that are creative successes will facilitate continued employment. Given this industry structure, how should a Hollywood professional navigate which projects to seek employment with, given only the basic information that would be available before a has a full cast and crew and begins production?

3 Data

This paper uses a dataset from Kaggle which includes movie metadata compiled from GroupLens and TMDB. The dataset includes metadata for each of 45,000 movies released during or before July 2017. In this case, metadata is basic information about the film, such as its budget, which series it belongs to (e.g. a Bond movie or Toy Story), applicable genres, and more. The dataset also includes quality ratings (and number of ratings) for each movie from 270,000 users for a total of 26 million individual ratings. Ratings range from 1-10. For a complete list of variables, see Appendix A.

The data require minimal cleaning: certain unneeded variables (such as links to images of movie posters) were removed; character variables were cleaned by removing punctuation, additional words (such as "name:" preceding each genre) and their numeric identifiers so that only the name of each genre remained. I converted the released date from a factor variable to a Date object and converted budget and revenue to numeric values. I also removed movies that had missing rating or revenue information or had budget values of 0 (which might indicate missing budget information).

Finally, I filtered the dataset to include only a subset of movies that could be considered relevant to the research question and avoid too many confounding variables. I wanted to

capture what would be considered "Hollywood" (or at least Hollywood-adjacent) films that would get a mainstream, theater release. To do so, I omitted movies that are considered "adult", movies not in English (some movies in English still have additional languages spoken in the film, but to be included, the movie had to have English as its primary language), those that were not produced in the United States (many of the movies were produced in additional countries), and movies that went straight to video. To ensure that the prediction algorithm can be tested against actual rating and revenue data, I eliminated movies that had not been released. Finally, I removed movies that were released before 2000 to avoid some confounding variables (e.g. more inflation, changing trends). After cleaning and subsetting, 4,127 movies remain in the dataset for analysis.

4 Empirical Methods

The primary goal of this paper is to determine which machine learning algorithm can most accurately predict movie outcome category. The target variable, therefore, is movie category. The FiveThirtyEight project that inspired this paper uses an intuitive, rather than machine learning, approach to movie classification (see Appendix B for an example of how they violate some typical rules of classification). Initially, I created four categories that were based simply on the mean values of each outcome variable (i.e. one category would be above-average rating, below-average revenue, etc.). A graph of the distribution of movies within these categories can be found in Figure 1. These are meant to be more "intuitive" categories, as in an average movie-viewer might make distinctions in this way (e.g. "that was a great movie that did poorly at the box office - I'd call it a cult classic"). Lash & Zhao

do something similar in splitting movies into three equally-sized groups based on ROI (Lash and Zhao, 2016).

In addition to these "intuitive" categories, I also use k-means clustering to classify the movies based on revenue and average rating. K-means clustering creates a specified number of categories while minimizing the within-cluster sum of squared Euclidean distances (WCSS) and can be represented as minimizing the following objective function:

$$J = \Sigma_{j=1}^{k} \Sigma_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \tag{1}$$

Where k is the number of clusters, n is the number of cases, x_i is case i, and c_j is the centroid of cluster j.

I find that the optimal number of classes is 4, which results in a within-cluster sum of squares/k of 88%. Decreasing k to 3 results in a drop of the WCSS/k to 81.9%, and increasing it to 5 leads to a WCSS/k of 92.1% (as well as a cluster of only 12 movies), so 4 seems to be the "elbow" where overfitting is avoided. I conduct two separate analyses, one using the "intuitive" categories as the target variable, the other using the k-means generated classes.

The prediction variables include the movie budget, a dummy variable for whether the movie belongs to a collection, a dummy variable for each of the eight major production companies (Ferriani et al., 2008), a dummy variable for each release season, a dummy variable for each genre, and a dummy variable for each of the top ten most frequent non-U.S. production countries in the data. The release season variable is imperfect because release may be rushed or delayed and so actual release season might be different from planned re-

lease season. However, it is included since most movies should be released around the time they were originally planned to be, although their planned and actual release dates may be different. For simplification, "release season" is based on month only.

For both analyses (using the "intuitive" categories and the k-means clustering classes as targets), I utilize five machine learning algorithms: decision trees, logistic regression, naive Bayes, k nearest neighbor, and support vector machine. The models are tuned using 6-fold cross-validation with 10 max iterations. Optimal parameters are obtained by minimizing mean missclassification error, a simple measure of the rate at which individuals are predicted to be in the incorrect category.

5 Research Findings

I find that none of the algorithms are strong when predicting the "intuitive" movie categories. The neural network algorithm is slightly more successful than the others, with a mean misclassfication error of 47.94%. This suggests that the "intuitive" categories may correspond to how the average moviegoer or even a movie professional thinks about classifying film success, but they were not constructed with much regard to the actual shape of the data and as a result are close to arbitrary. The relative success of the neural network model is not surprising given the use in previous literature of neural networks to successfully predict movie revenues using pre-release characteristics (Sharda and Delen, 2006). The ability of neural network algorithms to learn arbitrarily complex non-linear functions allows it to more accurately make classification predictions .

By contrast, when the classes generated through k-means clustering are the target vari-

able, all the algorithms (with the exception of naive Bayes) are much more successful at making predictions. In this analysis, the support vector machine algorithm is slightly more successful, with a mean misclassfication error of 22.94%. The SVM algorithm is able to easily handle the creation of arbitrary non-linear boundaries through the use of kernels, which likely helps its success in this case. In addition, SVM has an advantage over the neural network in this analysis because it operates on similar principles to k-means clustering (i.e. the construction of vectors and measuring of distances).

The extreme weakness of naive Bayes, which has a mean misclassification error of 91.91% in the second analysis, is likely related to its assumption that the input variables are independent. There are most likely strong dependencies within the input variables of the data. For example, certain genres frequently appear together. Using Bayes Theorem, the likelihood of a movie being in the "family" genre given the "animation" genre is 84% and the likelihood of a movie being a "drama" given the "romance" genre is 66%. Additionally, there is most likely covariance among the dummy variables for production companies, between production companies and budget, genres and budget, and budget and whether the movie belongs to a series. However, it is unclear why this algorithm is failing to such a great extent.

6 Conclusion

There is a significant difference in prediction success between the analysis using "intuitive" but non-data driven categories and classes generated by k-means clustering. In both analyses, all algorithms' performance metrics fall within a narrow range (with the exception of the poor performance of naive Bayes in the second analysis). Neural network narrowly beats the

other models in the first analysis, while support vector machine narrowly wins in the second.

The less arbitrary nature of the classes generated by k-means clustering makes prediction much more successful. However, since these classes are not intuitive, they would require additional interpretation to be used as decision-making tools for a movie professional trying to predict whether a movie they work for will be successful in the important dimensions of revenue and audience reception.

This paper has numerous weaknesses and limitations. First, movie ratings were based on a relatively small user base that might not reflect the general public's reception of a movie. Furthermore, there are additional pre-production movie characteristics for which data is hard to obtain or analyze that might substantially improve the ability to predict outcomes. For example, a movie that has not yet entered production might not have a cast or crew, but it most likely has at least a script which could be analyzed for tone and quality. At later and later stages of pre-production, more and more information (e.g. the reputation of the director or earliest actors to sign on to a film) becomes available to professionals considering employment on a particular project . Because this timeline may vary from film to film (e.g. a writer might have a particular actor in mind for a role, and so the cast starts to form before a director is hired), it would be hard to pick a consistent "point" in the movie timeline on which to base the analysis.

As the movie industry evolves with a greater emphasis on streaming rather than theater releases, further work is necessary to understand how movie success can be predicted. Future work could aim to assist investors in making decisions about which movies to fund or allow movie professionals (especially those just entering the industry) to make better-informed decisions about how to bolster their reputations.

References

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.

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.

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.

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

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.

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.

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

W. Hickey. The three types of sandra bullock movies, 2015a. URL https://fivethirtyeight.com/features/the-three-types-of-sandra-bullock-movies/.

W. Hickey. The four types of will ferrell movies, 2015b. URL https://fivethirtyeight.com/features/will-ferrell-movies-career-get-hard/.

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.

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.

S. Lumet. *Making Movies*. Vintage Books, New York, 1995.

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.

R. Sharda and D. Delen. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30:243–254, 2006.

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

7 Tables and Figures

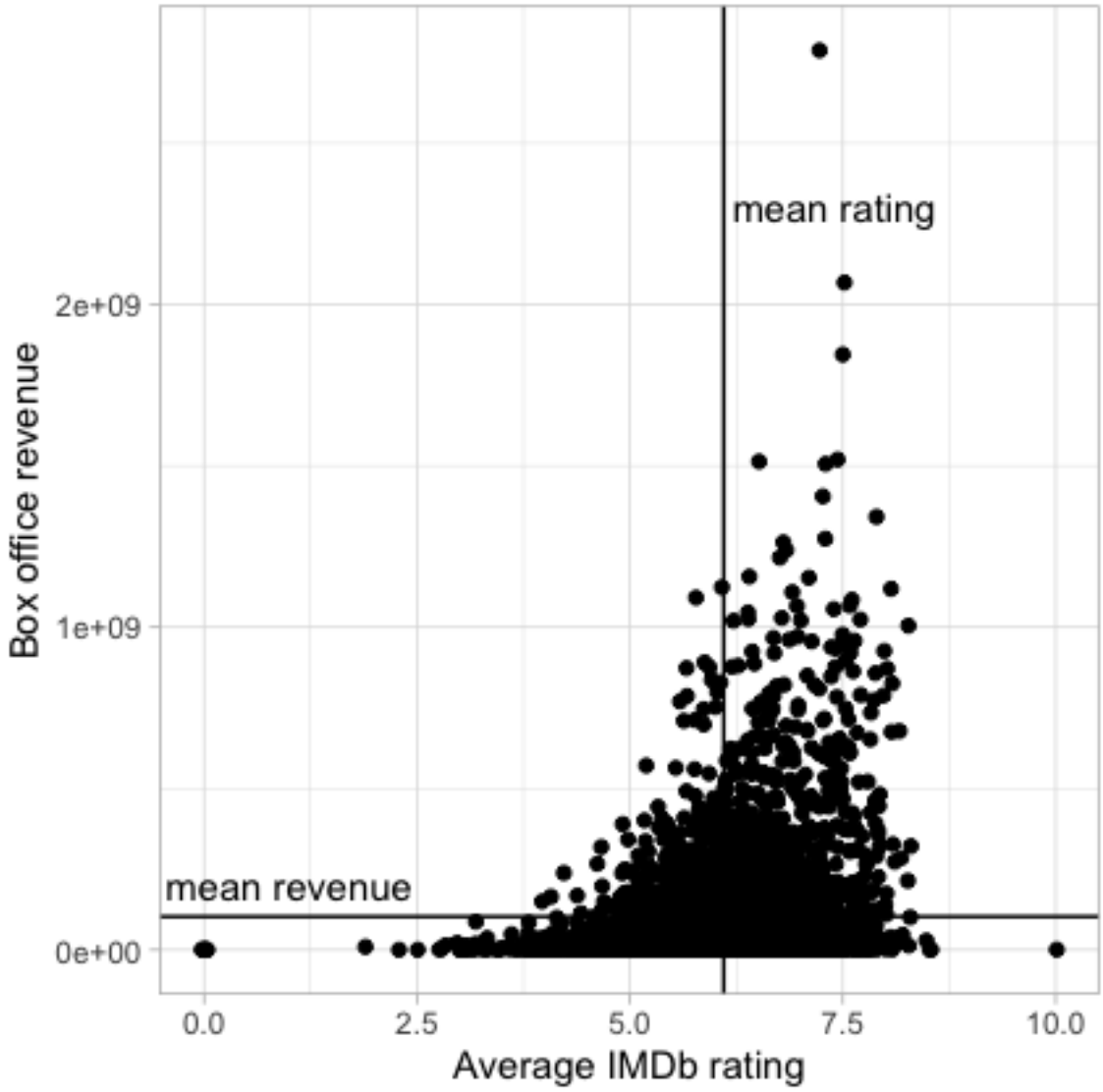

Figure 1: The distribution of movies by revenue and rating. The "mean rating" and "mean revenue" lines delineate movie categories.

Algorithm	Mean Misclassification Error
Decision Trees	0.5447942
Log Regression	0.5338983
Neural Network	0.4794189
Naive Bayes	0.5690073
kNN	0.4939467
SVM	0.5484262

Table 1: Mean misclassification errors for each prediction algorithm, using "intuitive" movie categories as the target variable.

Algorithm	Mean Misclassification Error
Decision Trees	0.2442244
Log Regression	0.2376238
Neural Network	0.2359736
Naive Bayes	0.9191419
kNN	0.2475248
SVM	0.2293729

Table 2: Mean misclassification errors for each prediction algorithm, using k-means clustering class as the target variable.

Appendices

A Movie Metadata 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