Colette Barca, Keith Osani, Nisha Srishan, and William Wulster

Professor Beecher

DATA 601-01

22 December 2020

Predicting a Film's Success Before Filming Begins

Business Context

In the film industry, "six or seven out of ten films are unprofitable" (usamnet000). Although the industry can be extremely risky, it has the potential to be highly profitable, "with major production houses raking in hundreds of millions of US dollars for specific projects" (Rodriguez). From 2016 – 2020, approximately 3,660 movies were produced annually (Nash Information Services). The box office and home entertainment sectors release films covering themes and topics that appeal to a wide range of consumers. Despite their variation in content, all films can be described using one set of characteristics. Every movie is associated with a genre, production company, and a team of individuals (actors, directors, producers, and writers). Even before production begins, all these components can be compiled into a succinct bundle, commonly referred to by industry professionals as "film packaging", which can then be used to garner financial support for the project (Schiller).

Although films can be extremely profitable, creating them is usually quite expensive. Typically, a movie's budget can range from $200,000 to $75,000,000, depending on the size of the project. These hefty price tags frequently necessitate "an elaborate patchwork of investors, banks, soft money tax credits and in-kind services" before filming can commence (Arnold). Due to the amount of risk involved, potential investor uses the information provided by film packaging to estimate his or her return on investment (ROI), determining if a particular movie will be

profitable. In our project, we attempted to create a statistical model for prospective financiers that would predict if a movie would be successful based on its proposed featured actors, director, producer, writer, genre, and production company.

Data

To build our model, we used the "IMDb Movies Extensive Dataset" on Kaggle (Leone). This dataset uses twenty-two features to describe 85,855 movies released between 1894 and 2020. Instead of using all these observations, we chose to perform our analysis on a subset of the data. Specifically, we restricted our dataset to only include films which were produced in the United States since 1980 with a budget of at least two million U.S. dollars. To be included, an observation also needed to contain a budget amount and worldwide gross income, both in terms of U.S. dollars. This resulted in a dataset consisting of 5,697 movies.

Since the term "success" is inherently ambiguous, we chose to approach this problem from the perspective of a potential investor and quantified a film's success based on its ROI. Determining a film's success using its percent ROI, rather than a straight ROI value, helped to normalize our comparisons. This allowed us to ignore variations in budget size and inflation rates. By definition, the ROI calculates the proportion of net income to the investment's cost (Beattie). To calculate the percent ROI for each movie, we applied the formula shown in Figure 1. We then worked together to determine an appropriate percent threshold to signify a successful movie. We could not find any literature that definitively evaluates a particular ROI percentage as successful. Hence, we chose to explore multiple percentages in an attempt to determine it ourselves. Ultimately, we decided to analyze three ROI percentages: 0%, 55%, and 100%. Technically, a movie that breaks even could be considered successful, as it would not cause investors to lose

$$\frac{worldwide\ gross\ income - budget}{budget} * 100$$

Figure 1: Movie Percent Return on Investment Formula

their money; however, they would walk away from the project with no financial gains. Meanwhile, a movie that yields a one-hundred percent return indicates a financier matched his or her investment. Since these two percentages signify opposite ends of the success spectrum, we felt it would be interesting to compare their results. Finally, we selected 55% because it almost perfectly split the number of successful and unsuccessful films in our dataset. Table 1 displays the number of successful and unsuccessful movies, respectively, for each ROI percentage. We used these three ROIs to create three distinct csv files, one for each percentage. In each csv, we added a dummy variable for the success flag; a value of one indicates the movie was successful, and a value of zero indicates the movie was unsuccessful, based on the corresponding ROI.

Table 1: Number of Successful vs Unsuccessful Movies by ROI

Return on Investment	Number of Successful Movies	Number of Unsuccessful Movies	Successful vs. Unsuccessful
ROI ≥ 0%	3,531	2,166	62% vs. 38%
ROI ≥ 55%	2,852	2,845	50% vs. 50%
ROI ≥ 100%	2,418	3,279	42% vs. 58%

With clearly defined parameters for a “successful movie”, we proceeded to develop a formula to transform the predictors’ textual data into numerical values. In our analysis, we chose to focus on six of the features: a movie’s genre, director, writer, producer, production company, and featured actors. We felt these six would be the most appropriate for our modeling, as they are the pieces of information that would be provided to investors pre-production. For each movie, we wanted to compute weight quantities for every predictor. We used two methods to accomplish this, based on the attributes of the feature.

For the columns that corresponded to individuals, we wanted to use each entry’s (i.e. each director’s, writer’s, actor’s, and producer's) previous successes to determine his or her respective weight on a film. Initially, we chose to simply divide an observation’s number of successful

movies by its total number of movies. However, with this method, an actor who starred in three successful films over his four featured movie appearances was rated as successful as an actress who starred in sixteen successful films over her twenty featured movie appearances. To avoid inadvertently assigning entries with very little experience identical scores to those with a great deal of experience, we chose to normalize the success weights. We replaced the observed values with their respective z-scores, allowing us to view each observation's predictor-success rate compared to the dataset's average predictor-success rate (Glen). Although normalizing the success rates allowed us to fairly compare all observations, our weights were still inaccurate. This method assumed weights were constant for all of an artist's films in each dataset. Since we approached this project from a potential investor's perspective, we needed to consider how information would be presented to financiers during a pitch. Due to "the risk involved and the desired return on investment," investors tend to prefer movies with components "that have a proven track record of success at the box office" (Arnold). Furthermore, an individual's level of success can grow and decline over time. To ensure our weights reflected this, we chose to compute a cumulative normalized success rate. We will illustrate this process below, using the 1995 film *Toy Story* and its principal actor, Tom Hanks, as an example. To compute Tom Hanks's weight on the film we would calculate the following:

$$\frac{\frac{\textit{Number of Successful Films Tom Hanks was in Before Toy Story}}{\textit{Total Number of Films Tom Hanks was in Before Toy Story}} - \textit{Average \% Actor Success of } 1994^{\bigstar}}{\textit{Standard Deviation of \% Actor Success of } 1994^{\bigstar\bigstar}}$$

$^{\bigstar}$Average % Actor Success from 1980 – 1994
$^{\bigstar\bigstar}$Average Standard Deviation of % Actor Success from 1980 – 1994

We would then repeat this process for the remaining featured actors in *Toy Story* (Tim Allen, Don Rickles, and Jim Varney), followed by taking the average of these four weights. This resulting value would be *Toy Story*'s overall actor weight. This process would be repeated for the

three remaining artist predictors. Notice that this equation is contingent on an artist's prior experience as well as the year the film was released.

We took a similar cumulative approach with the genres and production companies; however, we did not feel it was necessary to incorporate the average and standard deviation of all genre successes. Unlike an individual where we wanted to take newcomers into consideration, there are not too many new genres or production companies each year. Returning to the *Toy Story* example, to compute the Animation, Adventure, Comedy genre's weight on the film, we would calculate the following:

$$\frac{Number\ of\ Successful\ Films\ with\ the\ Animation, Adventure, Comedy\ Genre\ Before\ Toy\ Story}{Total\ Number\ of\ Films\ with\ the\ Animation, Adventure, Comedy\ Genre\ Before\ Toy\ Story}$$

This process would then be repeated for the movie's production company. After computing these six weights for each movie, we found there were 1,641 observations missing values for at least one of the artist weights, as the Kaggle dataset is missing some of the pertinent information regarding the individuals' prior film experience. Since there was no accurate way to input these values, we decided to remove these instances from our datasets, leaving us with 4,056 films.

Before starting our analysis, all samples from the years 2019 and 2020 were set aside from the dataset and reserved for model deployment. This subset contained 113 films. We decided to set a seed prior to randomly splitting the remaining data into training and testing sets. This ensured the four of us used the same exact observations to fit our models and assess their accuracies. Ergo, we could objectively compare our resulting accuracy percentages without having to account for sampling variation.We chose to proceed with the statistically common partitioning ratio of 2:1 observations for training data to testing data. Our training and testing sets contained 2,641 and 1,302 recruits, respectively.

Modeling

To model the data, we used two different techniques: Classification Trees and Logistic Regression. Since our data represent a classification problem, we decided to use two techniques to model our datasets and compare the results from both models. Each model was fit on the training data and then applied to the testing data to determine the model's accuracy. When fitting a model, we tuned its parameters to produce the highest possible testing accuracy. For the Classification Trees, we tuned the `max_depth` parameter in order to determine the depth of the tree that yielded the best model. When fitting with Logistic Regression, we created two different models for each dataset. The first Logistic Regression model was fit using all the predictors. Afterwards, we used the `summary2()` method from the `statsmodels` library to create a summary of the *p*-values for each predictor. We then refit the model using only the statistically significant predictors; specifically, those with a *p*-value less than 0.05. The specifications for the Classification Tree and Logistic Regression models fit on each dataset can be seen in Table 2.

Model Deployment

Table 2: Specifications of the Models Fit for Each Dataset

Percent ROI	Classification Tree Max Depth	Logistic Regression Significant Predictors
0%	2	actor_weight, writer_weight, producer_weight, production_co_weight
55%	4	All six predictors were statistically significant
100%	4	actor_weight, writer_weight, producer_weight, genre_weight, production_co_weight

After creating our models, we selected the best modeling technique, determined by the highest test score. In all three datasets, the Logistic Regression model fit with all the predictors produced the highest test accuracy. However, Table 3 reveals this technique just slightly outperformed the Classification Trees, differing by no more than half a percentage. In fact, when the ROI percentage was zero,

the two techniques produced identical results. Due to its slightly better performance, we used Logistic Regression to refit the model on the union of the testing and training data. For the zero percent dataset, we also refit using Classification Trees to see if one outperformed the other in model deployment. In all cases, we applied the new models to the hold out dataset, which included movies from 2019 and 2020. Each Logistic Regression model was fit once using all the predictors and then again with only the statistically significant predictors to determine which was more accurate.

Table 3: Testing Scores by Model and Dataset

Percent ROI	Classification Tree	Logistic Regression	Logistic Regression with Significant Predictors	Hold Out Data
0%	71.8126	71.8126	63.0321	75.2212 *Logistic Regression with the Four Significant Predictors
55%	66.6667	67.1275	All predictors were statistically significant	66.3717 *Logistic Regression with All the Predictors
100%	65.7450	65.8218	65.2726	65.4867 *Logistic Regression with All the Predictors

As shown in Table 3, the zero percent ROI's Logistic Regression model with only the four statistically significant predictors produced the most accurate predictions. In fact, this model was almost ten percent more accurate than the other best models. This suggests that, using the information typically presented with film packaging, it is possible to predict if a particular film will lose money. However, such information is not as effective for predicting how successful the project will be. Therefore, a different approach would need to be employed to help an investor identify the movie that would yield the largest profit.

References

Arnold, Kathryn. "The Basics of Film Finance." *HGExperts.com,* https://www.hgexperts.com/expert-witness-articles/the-basics-of-film-finance-7309. Accessed 13 Dec. 2020.

Beattie, Andrew. "A Guide to Calculating Return on Investment (ROI)." *Investopedia,* 31 Aug. 2020, https://www.investopedia.com/articles/basics/10/guide-to-calculating-roi.asp#:~:text=ROI%20is%20calculated%20by%20subtracting,finally%2C%20multiplying%20it%20by%20100.

Bell, Angelo. "Attract Film Investors." *FilmProposals,* https://www.filmproposals.com/Attract-Film-Investors.html. Accessed 12 Dec. 2020.

Dams, Tim. "AI Commissioning: Can Machines Aid Creativity in Filmmaking?" *IBC,* 19 Aug. 2020, https://www.ibc.org/features/is-the-film-industry-turning-on-to-artificial-intelligence/6417.article.

---. "Is the Film Industry Turning on to Artificial Intelligence?" *IBC,* 19 Aug. 2020, https://www.ibc.org/features/is-the-film-industry-turning-on-to-artificial-intelligence/6417.article.

Escandon, Rosa. "The Film Industry Made a Record-Breaking $100 Billion Last Year." *Forbes,* 12 Mar. 2020, https://www.forbes.com/sites/rosaescandon/2020/03/12/the-film-industry-made-a-record-breaking-100-billion-last-year/?sh=3008c05a34cd.

Frey, Bruno. *Economics of Art and Culture*. Springer International Publishing, 2019. *Google Books,* https://www.google.com/books/edition/Economics_of_Art_and_Culture/e7SPDwAAQBAJ.

Garon, Jon. "Film Financing: Equity and Debt Financing. *Gallagher, Callahan, & Gartrell,* Aug. 2009, http://www.gcglaw.com/resources/entertainment/film_financing.html.

Glen, Stephanie. "Z-Score: Definition, Formula and Calculation." *Statistics How To: Elementary Statistics for the Rest of Us!,* https://www.statisticshowto.com/probability-and-statistics/z-score/. Accessed 13 Dec. 2020.

gwmfilms (Maria Johnsen). "The End of Agency Film Packaging." *Golden Way Media Films,* 3 Sep. 2020, https://www.goldenwaymediafilms.com/agency-film-packaging/.

Leone, Stefano. (2020, September). *IMDb movies extensive dataset*. Retrieved from Kaggle website: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset.

Meslow, Scott. "How Hollywood Chooses Scripts: The Insider List That Led to 'Abduction'." *The Atlantic,* 23 Sep. 2011, https://www.theatlantic.com/entertainment/archive/2011/09/how-hollywood-chooses-scripts-the-insider-list-that-led-to-abduction/245541/.

Nash Information Services. "Movie Index." *The Numbers,* https://www.the-numbers.com/movies/#tab=year. Accessed 19 Dec. 2020.

Pearson, Erin. "The Basics to Making a Low Budget Film." *Topsheet,* 27 Sep. 2019, https://topsheet.io/blog/basics-to-making-a-low-budget-film.

Rodriguez, Jack. "6 Highest Earning Film Production Companies." *Beyond the Sight,* https://www.beyondthesight.com/film-production-companies/. Accessed 19 Dec. 2020.

Schiller, Christopher. "It Depends: What is Film Packaging?" *Script,* 9 Feb. 2018, https://scriptmag.com/features/it-depends-what-is-film-packaging.

"The History of Movies." *Understanding Media and Culture: An Introduction to Mass Communication*. Saylor Academy, 2012. https://saylordotorg.github.io/text_understanding-media-and-culture-an-introduction-to-mass-communication/s11-01-the-history-of-movies.html.

usamnet000. "Reasons for Success and Failure in the Movie Industry." *DEV,* 19 Jun. 2020, https://dev.to/usamnet000/reasons-for-success-and-failure-in-the-movie-industry-25m2.

Vogely, Harold. *Entertainment Industry Economics: A Guide for Financial Analysis*. Cambridge U.P., 2019. *Google Books,* https://www.google.com/books/edition/Entertainment_Industry_Economics/zjHg5j0CsEoC.

Watson, Amy. "Box Office Revenue in North America from 1980 to 2019." *Statista,* 10 Jan. 2020, https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/.