Early Prediction of Movie Success — What, Who, and When

Michael Lash[1(✉)], Sunyang Fu[2], Shiyao Wang[1], and Kang Zhao[2(✉)]

[1] Department of Computer Science, The University of Iowa, Iowa City, USA
{Michael-Lash,Shiyao-Wang}@uiowa.edu
[2] Tippie College of Business, The University of Iowa, Iowa City, USA
{Sunyang-Fu,Kang-Zhao}@uiowa.edu

Abstract. Leveraging historical data from the movie industry, this study built a predictive model for movie success, deviating from past studies by predicting profit (as opposed to revenue) at early stages of production (as opposed to just prior to release) to increase investor certainty. Our work derived several groups of novel features for each movie, based on the cast and collaboration network (who'), content ('what'), and time of release ('when').

Keywords: Predictive model · Data analytics · Social network · Text mining · Decision support system

1 Introduction

In the U.S., the motion picture industry produces approximately 500 movies in a year [10] garnering, on average, $60 million of investment capital per film [12]. Despite the large capital investment that must be made prior to movie production, the success, or profitability, of a movie is largely uncertain. Past studies have two limitations: First, the focus is almost solely on total box office revenue [2], [9], [11] or theater admissions [8]; Second, many of these studies employed features that are only available just prior to [1,2], or even after [7] [12] [14], the official release of a movie.

For investors, though, the profit of a movie is a better indicator of success than the total revenue. Also, it is worth noting that the prediction of a movie's success should leverage features that are available only during the time in which investments are being garnished.

To the best of our knowledge, this research represents the first attempt to predict the profitability of movies at early stages of production, which we achieve by examining 'who' factors-- its actors, directors, and social networks based on previous collaboration; the 'what' factors—a movie's genre, rating, and plot synopsis; as well as the 'when' factor—when a movie will be released and the temporal ebbs and flows of the movie industry.

N. Agarwal et al. (Eds.): SBP 2015, LNCS 9021, pp. 345–349, 2015.
DOI: 10.1007/978-3-319-16268-3_41

2 Feature Engineering

2.1 Who are Involved

The movie industry is characterized by movie stars, which function as a name brand, drawing crowds, and thus increasing sales [2], [6] . Thus we included the following features to measure the 'star power' of a movie.

a. Total and average tenure: Total/average between first and most recent appearance.

b. Total and average actor gross are the sum/average of all actors' gross revenues across all movies they have appeared in.

d. Total and average director gross are the total/average revenues of all movies directed by the director of target movie m.

Another important factor for movie success is team cohesion [8]. Thus we took a social network approach, which provide a wealth of valuable information about various types of inter-personal relationships, such as dating [15], collaboration [16], and communication [5]. In this research, we built a collaboration network among actors based on their co-star relationship; we aggregated this undirected, weighted (collaborations) network to 1999 and created 11 separate, yearly networks (through 2010); each year is used in deriving features to predict a subsequent years profit.

Network metrics that fall into the former category can be summarized as follows:

a. Team heterogeneity: It is believed that successful movies have a certain degree of originality to them [8], which is accomplished by a hetergenous team of actors, and which we capture via cosine similarity using cast members' neighborhood vector, denoted in Equation 1.

$$\frac{1}{(n(n-1)/2)}\sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1}\frac{Act_i \bullet Act_j}{\| Act_i \| \| Act_j \|} \tag{1}$$

b. Average degree: overall centrality among team members.

c. Total and average betweenness centrality: experienced actors who are not well known may bring an 'unseen' benefit to a production.

Network metrics that measure a movie's effect on the structure of an existing social network are summarized as follows:

d. Decrease in average shortest path: Structural holes are an important concept in networks pertaining to movies [13], as individuals who span these structural holes are said to have high social capital [4].

e. Change in clustering coefficient: A feature that captures the diffusion of information across the network [16].

We capture content-based features, such as those found in a script or, in our case, a plot synopsis, by using Latent Dirichlet Allocation (LDA) [3], which his ultimately expressed as a topic distribution vector for each movie in our dataset.

2.2 When a Movie will be Released

Temporal features are an important component to consider in the ever-evolving multi-billion dollar motion picture industry. As such, we incorporated the following features:

a. Average annual profit in the year prior to release.

b. Annual profitability percentage by genre is the percentage of profitable movies in the previous year, that fall into the same genre as *m*.

c. Annual weighted profitability by genre (AWPG): Is the sum of cosine similarity, genre vector-wise, between a given movie and each movie of the previous year, weighted by the profitability of that genre.

$$AWPG_m = \sum_{m_i' \in y-1} sim\big(g(m), g(m_i')\big) * p(m_i') \quad (2)$$

d. Release dates: *Season* (spring, summer, etc) and *Holiday Release* (the week leading up to, and including, Christmas).

3 Experiments and Results

Data was collected from a movie archive website, BoxOfficeMojo, including data of 5,440 actors as well as 14,097 movies from year 1921 to 2014.

3.1 Experiment Setup and Results

The goal of our research is to predict whether a movie will be profitable, which we define as a profit (revenue minus budget) at .25 of 1 standard deviation above mean, in order to insure a reasonable ROI.

Our dataset includes 1353 movies (revenue and budget info available), 384 of which were profitable. We excluded sequels (confounding) and features pertaining to cast members were derived from the full set of 14097 movies.

Table 1. Classification outcome from the logistic regression classifier

	All features	without When	without What	without Who
AUC	0.801	0.736	0.803	0.72
Accuracy	0.771	0.738	0.763	0.734
F1 score	0.757	0.714	0.743	0.708

Various classification algorithms were used for the prediction (with 10-fold cross validation) with logistic regression yielding the best results, (please refer to the 'All features' column in Table 1).

3.2 Discussions

Results in Table 1 also show those obtained from our logistic regression classifier upon omitting each class of features. As the reader will notice, the classifier deteriorates when

'When' and 'Who' features are removed, indicating the contribution of such features; 'What' features, then, likely overlap those of 'When' and 'Who'.

Table 2 shows the top contributing features by coefficient value of the "non-profitable" class (negative coefficient are indicative of profit), which admittedly only indicate their role within our predictive model. However, there are still some very interesting findings. For example, the Horror Genre contributes positively to profitability, perhaps due to the relatively low budget needed for success (ie. Paranormal Activity); "Documentaries" may be in a similar situation. It was also found that annual profit percentage by genre contributed to movie profitability, indicating the importance of current trends. In terms of "who" is involved in a movie, both average director gross and average actor gross contribute positively to movie profitability (name brand and skill).

Table 2. Top features in each feature group for non-profitable movies

Group	Feature	Coefficient
What--Rating	"NC-17" rating	12.13
	"G" rating	-1.047
What--Genre	"Documentary" genre	-1.609
	"Horror" genre	-1.297
What--Plot topic	Topic #13	0.343
	Topic #7	0.183
Who--Individual	Avg. director gross	-0.963
	Avg. actor gross	-0.906
Who--Network	Avg. degree	0.362
	Total betweenness centrality	0.159
When	Annual prof. pctg. by genre	-1.007
	Winter release	-0.297

On the other hand, some features were noteworthy with regard to movie losses; topic 13, represented by familial-type words (eg. "wife"), indicating that perhaps movies that focus their content on relationships are less often profitable. While the NC-17 rating is of no surprise, average degree and total betweenness centrality are; we believe this may be a product of collinearity, better address by an explanatory model in future work.

4 Conclusions and Limitations

In this paper, we predicted the profitability of movies at early stages of movie production, leveraging "What", "Who" and "When" features. Applying this predictive model to empirical data, we showed that our model is able to achieve decent performance, using a wide variety of features, which could be employed in decision support to aid in investment decisions

The authors recognize that there are a few limitations, including collinearity and possible sampling bias.

References

1. Apala, K.R., Jose, M., Motnam, S., Chan, C.-C., Liszka, K. J., de Gregorio, F.: 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, ASONAM 2013, pp. 1209–1214. ACM, New York (2013). doi:10.1145/2492517.2500232
2. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Washington, DC, USA, pp. 492–499 (2010). doi:10.1109/WI-IAT.2010.63
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). doi:10.1162/jmlr.2003.3.4-5.993
4. Burt, R.: Structural holes: The social structure of competition. Harvard Univ Press (1995)
5. Diesner, J., Frantz, T., Carley, K.: Communication Networks from the Enron Email Corpus 'It's Always About the People. Enron is no Different'. Computational & Mathematical Organization Theory **11**, 201–228 (2005). doi:10.1007/s10588-005-5377-0
6. Elberse, A.: The Power of Stars: Do Star Actors Drive the Success of Movies? Journal of Marketing **71**(4), 102–120 (2007). doi:10.2307/30164000
7. Gopinath, S., Chintagunta, P.K., Venkataraman, S.: Blogs, Advertising, and Local-Market Movie Box Office Performance. Management Science (2013). doi:10.1287/mnsc.2013.1732
8. Meiseberg, B., Ehrmann, T.: Diversity in teams and the success of cultural products. Journal of Cultural Economics **37**(1), 61–86 (2013). doi:10.1007/s10824-012-9173-7
9. Meiseberg, B., Ehrmann, T., Dormann, J.: We don't need another hero–implications from network structure and resource commitment for movie performance. Schmalenbach Business Review (sbr) **60**(1), 74–98 (2008)
10. Mestyán, M., Yasseri, T., Kertész, J.: Early prediction of movie box office success based on Wikipedia activity big data. PloS One **8**(8), e71226 (2013)
11. Sharda, R., Delen, D.: Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications **30**(2), 243–254 (2006). doi:10.1016/j.eswa.2005.07.018
12. Simonoff, J.S., Sparrow, I.R.: Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers. Chance **13**(3), 15–24 (2000). doi:10.1080/09332480.2000.10542216
13. Zaheer, A., Soda, G.: Network Evolution: The Origins of Structural Holes. Administrative Science Quarterly **54**(1), 1–31 (2009). doi:10.2189/asqu.2009.54.1.1
14. Zhang, W., Skiena, S.: Improving movie gross prediction through news analysis. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Washington, DC, USA, pp. 301–304 (2009). doi:10.1109/WI-IAT.2009.53
15. Zhao, K., Wang, X., Yu, M., Gao, B.: User recommendation in reciprocal and bipartite social networks–an online dating case study. IEEE Intelligent Systems **29**(2), 27–35 (2013). doi:10.1109/MIS.2013.104
16. Zhao, K., Yen, J., Ngamassi, L.-M., Maitland, C., Tapia, A.: Simulating Inter-organizational Collaboration Network: a Multi-relational and Event-based Approach. Simulation **88**, 617–631 (2012). doi:10.1177/0037549711421942