How to Make a Successful Movie: Factor Analysis from both Financial and Critical Perspectives

Zheng Gao[1(✉)], Vincent Malic[1], Shutian Ma[2], and Patrick Shih[1(✉)]

[1] Indiana University Bloomington, Bloomington, USA
{gao27,vmalic,patshih}@indiana.edu
[2] Nanjing University of Science and Technology, Nanjing, China
mashutian0608@hotmail.com

Abstract. Over the past twenty years, people have seen considerable growth in film industry. There are two common measurements for movie quality, financial metric of net profit and reception metric in the form of ratings assigned by moviegoers on websites. Researchers have utilized these two metrics to build models for movie success prediction separately, while few of them investigate the combination. Therefore, in this paper, we analyze movie success from perspectives of financial and critical metrics in tandem. Here, optimal success is defined as a film that is both profitable and highly acclaimed, while its worst outcome involves financial loss and critical panning at the same time. Salient features that are salient to both financial and critical outcomes are identified in an attempt to uncover what makes a "good" movie "good" and a "bad" one "bad" as well as explain common phenomenons in movie industry quantitatively.

Keywords: Movie success prediction · Social network analysis · Feature construction

1 Introduction

These days, people are deeply influenced by the film industry from both financial and cultural aspects. According to the Box Office Mojo annual report[1], 724 movies were released in 2017 and this industry generated over $10 billion gross in the United States domestic market alone. Such statistics are just the latest indication of a consistent growth in both number of movies produced and amount of money earned over the past 20 years. With the film industry firmly positioned as a pillar of cultural production in the 21st century, the question of what makes a movie financially and critically successful is worthy of investigating.

As we can see, most researchers suppose that critical success or financial success of the movie can represent its overall success directly. However, we find

[1] Box office mojo annual report: http://www.boxofficemojo.com/yearly/.

N. G. Taylor et al. (Eds.): iConference 2019, LNCS 11420, pp. 669–678, 2019.
https://doi.org/10.1007/978-3-030-15742-5_63

that, these two kinds of successes are not correlated with each other. Ideas of investing more money to obtain better critical reception, or crafting a well-designed movie to aim for a significant profit, are not necessarily valid. By way of illustration, the 2008 crime drama *Nothing But the Truth* has an average user rating of 7.2 with 31,490 votes, putting it above the 75^{th} percentile in terms of user reception. It made a total gross profit of $3,045 in the US, a staggering loss in light of the movie's $11,500,000 budget - essentially a complete failure in terms of ROI. The 2015 horror film *The Gallows*, on the other hand, has a user rating of 4.2 with 14,983 votes, which is below the 5^{th} percentile. It nevertheless obtained a total gross of $22 million dollars on a budget of $100,000 - an ROI of 226.58. In light of our primary findings and existing examples, it's argued that modeling financial and critical success simultaneously is distinct from modeling them separately. Therefore, in this paper, we predict movie success from critical and financial aspects at the same time.

The Contribution of This Work is Fourfold. Firstly, a combination of return on investment and user rating is defined as a composite criteria to evaluate movie success. Secondly, an quantitative analysis is conducted on not only basic features from metadata but also complex movie features calculated synthetically to determine the role these features play in light of our new success metric. Thirdly, the identified features are utilized in machine learning models to see if they are able to predict success of a given film. Finally, this paper is able to reveal three phenomenons which also exist in real movie industry:

1. Among all the genres of movie, family dramas tend to attract audience more easily;
2. The success of a movie heavily relies on the success of its cast's past career.
3. Stable collaboration between directors and actors are more likely to achieve long term movie success especially in series movies.

2 Related Works

The motion picture industry in the United States is a big business. A report by the industry tracking firm Nash Information Services shows that ticket sales have grown steadily over the past 20 years[2]. This growth coincides with an increasing amount of data about movies, which researchers have turned to in order to find ways to discover features that characterize blockbusters or flops [17] and to examine the interplay between ratings and revenue after a movie has been released [14].

Various kinds of social media platforms are heavily involved in gross earning predictions. some researches use features generated from those online open resources to predict gross earnings. In [1], Armstrong and Yoon extracted features from IMDB and used a regression model to predict the user rating of a movie, while other works focus specifically on the effect that the "star power"

[2] http://www.the-numbers.com/market/.

of the leading actors and actresses has on a movie's reception and income [8]. In [12], the authors rely on Twitter buzz surrounding a film's release to predict its box office revenue. [11] also extracts data from Twitter, but measure the sentiment present in the Twitter conversation to see if such sentiment effects box office performance. In [13], the authors look at Wikipedia activity surrounding upcoming films to make similar predictions. Other large scale social medias are also discussed in [6] to explore its power to movie success.

Some researches argue that user reviews are not a helpful indicator for predicting box office revenue [5]. However, such reviews still remain a factor contributing to success [3] since many potential audience members refer to such reviews when deciding to see a film. Previous research studies explicitly show the influence that user ratings and individual reviews have on prospective audience members [19]. Existing works [15] use movie reviews on Twitter for profit prediction. Some other works such as [10] examine critical reviews from other review sites to see if such sites are predictive of revenue. Moreover, besides online review, there are other ways to spread movie information such as news and word of mouth, etc. Those methods are also used for movie profit prediction [20] to see how much effect rating can influence movie profit. While earlier research dealt with primary features extracted from movie datasets or social medias, another research focuses on generating novel features from existing ones to improve prediction performance [18]. Lash and Zhao [9] obtain new gains in predicting movie ROI by systematically categorizing the types of features available for analysis and recombining them in novel ways.

3 Experiment

3.1 Dataset

Our data consists of data from IMDB on movies produced before October 2016. As the movie industry boosts in recent 20 years based on motion picture yearly report, movies released domestically within this period are the valid examples to use for exploring the reason. After keeping movies released in past 20 years with abundant meta information, 6,981 movies remain.

The average IMDB user rating is used as a metric for critical performance and represent financial performance through Return on Investment, which is defined as:

$$ROI = \frac{gross - budget}{budget} \tag{1}$$

A movie that makes a large gross profit is not necessarily a "profitable" movie. A general rule of thumb for qualifying a movie as a "financial success" is to compare its gross revenue to twice its reported budget - in other words, an ROI of at least 1. A movie "doubly" is successful if it performs well both financial and critically while deem it a "total" failure if it loses money and is panned. Therefore a label of "success" is assigned to a movie that attains ROI bigger than 1 and its average user rating is above the global user rating average, and a

label of "failure" means a movie's ROI is less than 0 and user rating is below the global average. Under this schema, 2,076 movies qualify as successes and 1,960 are failures.

3.2 Feature Calculation

To incorporate elements of the movie's plot into our model, Latent Direchlet Allocation [4] is utilized to create a topic model of movie plot summaries, which can generate a series of ranked topics associated with a list of ranked words in each topic to quantitatively represent movie plots in a latent vector space. [7] uses Gibbs sampling method on LDA to choose 15 optimized topics as the best representations for movie plots. We manual-coded the interpretation of the 15 topics based on the top ranked words, which are 'world war', 'misfortune', 'youths in trouble', 'crime drama', 'family', 'place stories', 'love and marriage', 'school life', 'survival', 'sex and relationships', 'marriage and family', 'making it in society', 'show business', 'political intrigue', 'high life drama'.

In addition to basic features belonging intrinsically to the movie itself, it is necessary to explore whether career performance of actors and directors as well as their collaborations affect movie success. Hence, actor and director data are aggregated for a given movie into a set of composite features. For example, the historical performance of actors in a film is calculated by finding the average user rating of all previous films for every actor in the current film, and then averaging those results to create the feature *average_ActorRating_average*. In the end Each movie is represented as a set of 24 features in three categories. Details are shown in Table 1.

3.3 Main Approach

Feature Correlation Detection. Our analysis first find and examine correlations in the obtained data. First, as all advanced features are synthetic, there is the potential that the synthetic features contain too high a degree of overlapping information. Second, it also offers a logical way to filter out some features containing duplicated information. Pairwise Pearson Correlation [2] is applied to detect correlation between all factor pairs. In this paper, if a pair of features meet both two criteria: (1) the correlation score between them is above 0.5 with a significant p value below 0.1; (2) their correlation similarity difference with the rest of features are all smaller than 0.1, one of the features will be removed.

Latent Feature Exploration. After feature construction and processing, two concerns still remain. First, as most of the features are high number of synthetic features may introduce a large amount of noise that any potential model may overfit on. Second, we also want to explore the possibility that the features we have arrived at are instead representations of latent variables which themselves are more influential in predicting movie success. Therefore, Principle Component Analysis (PCA) [16] is applied to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables. In the end, a

Table 1. Feature description

Feature name			Abbr.	Feature description
Basic features	Content based	genre	genre	movie genres (24 types)
		MPPA rating	rating	MPPA ratings (23 types)
	Time based	released year	year	the released year
		released season	season	the released season
		week day	day	the released week day
Advanced features	Actor based	average_ActorTenure	aAT	the average years of the actors career length
		total_ActorTenure	tAT	the total years of the actors career length
		total_ActorGross_total	tAGt	sum of all actors' career movie gross
		total_ActorGross_average	tAGa	sum of all actors' average movie gross
		average_ActorGross_average	aAGa	average of all actors' average movie gross
		total_ActorProfit_total	tAPt	sum of all actors' career movie profit
		total_ActorProfit_average	tAPa	sum of all actors' average movie profit
		average_ActorProfit_average	aAPa	average of all actors' average movie profit
		top_ActorProfit_average	tpAPa	the largest average value among all actors' average movie profit
		top_ActorProfit_top	tpAPt	the largest value of all actors' most profitable movie
		average_ActorRating_average	aARa	average of all actors' average movie rating
	Dirctor based	average_DirectorRating_average	aDRa	average of all directors' average movie rating
		total_DirectorGross_total	tDGt	sum of all directors' career movie gross
		total_DirectorGross_average	tDGa	sum of all directors' average movie gross
		average_DirectorGross_average	aDGa	average of all directors' average movie gross
	Collaboration based	actorDirectorCollab_frequency	aDCf	the number of times the actors collaborated with the directors before
		actorDirectorCollab_rating	aDCr	the average rating of all the movies that the actors and directors collaborated before
		actorDirectorCollab_profit	aDCp	the average net profit of all the movies that the actors and directors collaborated before
Topic modeling feature			topic	movie plot distribution on 15 topics

denser, lower-dimensional representation of the data is obtained which reveals latent relationships among features and reduces computation workload. Usually, components with eigenvalue above 1 contain noticeable feature information.

Prediction Model. Support Vector Machines (SVM) is one of the most widely used classification methods which uses a kernel function to separate instances

into different categories with good performance. In this paper, a Support Vector Machine with a linear kernel is trained to predict the outcome variable of financial and critical success. To evaluate model performance, 10-fold cross validation is leveraged to ensure the weight accuracy of each feature.

4 Results

4.1 Correlation Analysis

Correlation between advanced features as well as ROI and rating is presented as a heatmap in Fig. 1[3]. Red and blue indicate positive and negative correlation, respectively, while color saturation represents the magnitude of the correlation.

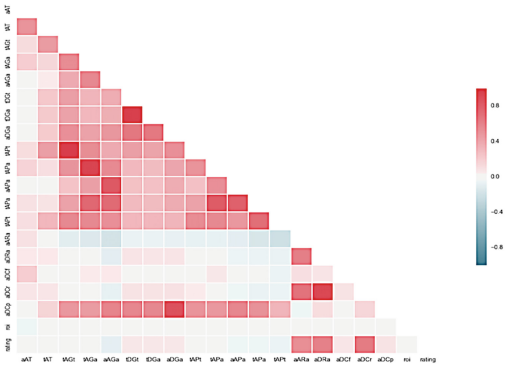

Fig. 1. Advanced features & ROI & rating correlation heat map (Color figure online)

The heatmap shows that most advanced features are not highly correlated, and ROI is not correlated with user rating. Quantitatively, from the correlation result of each feature associated with p value, there is no feature that meets the criteria to be eliminated. All features have the potential to contribute to movie success prediction.

One interesting finding is that features involving the history of gross box office for a given film tend to have a weak positive correlation among themselves and no correlation to ratings-based features. It implies a win-win collaboration between actors and directors. If their collaborations can help movies achieve more box office, they are willing to continue their relationship in future movies, which refers to common phenomenon (3) mentioned in the Introduction section. Basically for all American series movies such as *Transformers*, the directors and main actors keep same for all series as it can lead to huge box office. However not all movies in the whole series have high IMDB user rating. Another finding also implies common phenomenon (3): One of the highest correlations is between the

[3] Please zoom in if the font size is too small to read.

feature actorDirectorCollab_rating and average_DirectorRating_average. Both of these features are generated from the director's rating history, and the strong correlation here reflects the tendency for directors to work multiple times with a chosen set of actors if their movies achieve critical success.

4.2 Feature Impact

By applying Principle Component Analysis all the features are reduced to 5 components with an associated eigenvalue greater than 1 that are able to explain 66.62% of the variance in the success label. The amount of explained variance refers to how well the features can explain the difference between movie failure and movie success. Table 2 shows the result of PCA component matrix where larger weights implies more contribution to form the condensed component. The weighting of each feature in a particular principle component is shown only if it exceeds a threshold of 0.3.

Table 2. PCA component & SVM weight matrix

Feature		PCA components				
		PC-1	PC-2	PC-3	PC-4	PC-5
Basic features	genre (Drama)			0.9314		
	genre (Comedy)			0.5852		
	genre (Western)					0.8259
	rating (USA:TV-MA)				0.4208	
	rating (USA:PG)				0.3463	
Advanced features	tAPt	0.3262				
	aAPa	0.3905				
	tAGt	0.3794				
	tAGa	0.3344				
	tpAPt	0.3165				
	tpAPa	0.3097				
	aDCP	0.4477	0.3429			
	aDRa		0.3736			
	aARa		0.3467			
	aDGa		0.4099			
Topic modeling features	topic_5 (family)					0.5750
	topic_11 (marriage and family)					0.4356
Eigenvalue		6.7552	3.2879	2.8152	1.1912	1.0574
Explained variance		29.78%	14.49%	12.41%	5.25%	4.63%
SVM model weight		0.1704	1.2955	0.2243	−0.0262	−0.1447
Total explained variance		66.62%				

The first component PC-1 is composed of a mix of actor financial metrics and explains 29.87% of the variance in the label. PC-1 can be regarded as a composite actor feature based on a suite of financial indicators, which implies common phenomenon (2) listed previously: Actors make the most contributions for a movie success. Director features are weighted more strongly, and the analysis shows that when the actor and director financial features are aggregated, they are

capable of explaining a greater proportion of movie success, which also explains common phenomenon (3): stable collaboration is important.

Actor and director critical performance, in contrast, are part of the make up of the second principle component, along with the average profit of actor-director collaboration and the average gross earnings of the director. This principle component PC-2 explains 14.49% of the variance. The top 2 main components are basically formed by advanced features, which means there are existing latent factors of the actor-director relationship and the career histories of the actor and director influence movie success the most.

Component PC-3 combines and provides high weights to the comedy and drama genres, while component PC-4 accounts for the effect of certification. Those two components have really clear meanings. In component PC-3, genre "Drama" has a really high positive impact weight. That may be part of the reason why movie industry is willing to produce movies in this genre, which refers to the common phenomenon (1). Though PC-4 is clearly connected to certification, it is difficult to provide it with a consistent interpretation.

Finally, the interpretability of the retrieved principle components drops at principle component five, which explains only 4.63% of the variance. Here, we begin to see the effects of specific topics, as well as the Western genre. Two topics about families form the component PC-5, meaning that the most important movie topic to audience is "family", which also refers to common phenomenon (1).

4.3 Predictions

The original data is re-expressed in the form of the five principle components as the input to a Support Vector Machine model with a linear kernel to predict if a movie will be a success or failure in terms of financial and critical reception. Table 2 shows that for the prediction of success the SVM assigned a significantly higher weight to PC-2, once again suggesting that directors and rating history play a more significant role in the final success or failure of a given movie. In other words, actor financial performance explains the variance in success or failure, which refers to common phenomenon (2), while director and actor critical performance are more pertinent to a movie actually becoming successful. It notes that actor-director collaboration profit feature exists with a high weight in both PC-1 and PC-2, suggesting that this advanced feature is relevant to both variance in outcome and likelihood of success, which explains common phenomenon (3).

Table 3. Evaluation metrics

Accuracy	0.7915	Precision	0.7919
Recall	0.7915	F1 score	0.7914
Hamming loss	0.2084	Matthews coefficient	0.5835

The performance of the SVM model is shown in Table 3. Since we took measures to ensure that the categories were equal in number, the baseline accuracy is 50%. Our model attains an accuracy of 79.15%, indicating that the feature selection and the PCA process has successfully zeroed in on the features of movies that are pertinent to success or failure. Other than accuracy, other evaluation metrics attain satisfactory levels. Precision reflects that movie successes can be identified correctly while high recall value means most of successes are retrieved in the model. F1 score integrates both precision and recall. The Hamming loss indicates that the model can predict movie success correctly with little error. A relatively high Matthews correlation coefficient also testifies to the performance of the model. The evaluation metrics support the reliability of our prediction model and interpretation analysis.

5 Conclusion

Movie user rating and profit are two separate but related factors in judging the ultimate performance of a movie. As a result, there exists much research that focuses on one or the other, but to the best of our knowledge this is the first work that attempts to predict financial and critical success simultaneously. According our analysis, three common phenomenons for movie industry is well explained quantitatively. There are more interesting findings explored in this approach. Advanced features with actor and director information play a considerable role in a movie's success. The created composite features add to the power of the model. And movie genre and plot are another two important features for successful movie production. Movies about humanity, family and comedy are better welcomed by people. And violence, horror and cult movies are tend not to be as successful.

In future work, we hope to improve this work by incorporating social media which has proven successful in recent literature as a predictor of financial success, but not of combined critical and financial success. We also wish to examine if the factors for success differ substantially in different cultural settings by using data from the Chinese equivalent of IMDB, Douban. Given the growth and impact of the motion picture industry, our research should be useful for those who desire to invest in a movie that not only earns a decent profit, but is also has a high cultural impact.

References

1. Armstrong, N., Yoon, K.: Movie rating prediction. Technical report. Citeseer (1995)
2. Benesty, J., Chen, Y., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol. 2, pp. 1–4. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00296-0_5
3. Berg, J., Raddick, M.J.: First you get the money, then you get the reviews, then you get the internet comments: a quantitative examination of the relationship between critics, viewers, and box office success. Q. Rev. Film Video **34**, 101–129 (2017)

4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Brown, A.L., Camerer, C.F., Lovallo, D.: To review or not to review? Limited strategic thinking at the movie box office. Am. Econ. J. Microecon. **4**(2), 1–26 (2012)
6. Ding, C., Cheng, H.K., Duan, Y., Jin, Y.: The power of the "like" button: the impact of social media on box office. Decis. Support Syst. **94**, 77–84 (2017)
7. Griffiths, T.: Gibbs sampling in the generative model of latent Dirichlet allocation (2002)
8. Karniouchina, E.V.: Impact of star and movie buzz on motion picture distribution and box office revenue. Int. J. Res. Mark. **28**(1), 62–74 (2011)
9. Lash, M., Fu, S., Wang, S., Zhao, K.: Early prediction of movie success — what, who, and when. In: Agarwal, N., Xu, K., Osgood, N. (eds.) SBP 2015. LNCS, vol. 9021, pp. 345–349. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16268-3_41
10. Legoux, R., Larocque, D., Laporte, S., Belmati, S., Boquet, T.: The effect of critical reviews on exhibitors' decisions: do reviews affect the survival of a movie on screen? Int. J. Res. Mark. **33**(2), 357–374 (2016)
11. Lehrer, S., Xie, T.: Box office buzz: does social media data steal the show from model uncertainty when forecasting for hollywood? Technical report, National Bureau of Economic Research (2016)
12. Liu, T., Ding, X., Chen, Y., Chen, H., Guo, M.: Predicting movie box-office revenues by exploiting large-scale social media content. Multimed. Tools Appl. **75**(3), 1509–1528 (2016)
13. 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)
14. Moon, S., Bergey, P.K., Iacobucci, D.: Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. J. Mark. **74**(1), 108–121 (2010)
15. Oh, C., Roumani, Y., Nwankpa, J.K., Hu, H.-F.: Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. Inf. Manag. (2016)
16. Pearson, K.: Liii on lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901)
17. Ravid, S.A.: J. Bus. **72**(4), 463–492 (1999)
18. Sharan, P.: Movie success predictor. Indian J. Appl. Res. **6**(6) (2016)
19. Wang, H., Guo, K.: The impact of online reviews on exhibitor behaviour: evidence from movie industry. Enterp. Inf. Syst., 1–17 (2016)
20. Zhang, F., Yang, Y.: The effect of internet word-of-mouth on experience product sales—an empirical study based on film online reviews. Int. J. Bus. Adm. **7**(2), 72 (2016)