**Exploration of Parameters affecting Box Office Revenue of Hollywood Movies**

**Abstract** The Hollywood movie industry is a business with a high profile, and a highly variable revenue stream. It’s not surprising that movie studios are intensely interested in predicting revenues from movies. In this paper, we explore the parameters affecting the revenue of Hollywood movies. We shed light on how box office revenue is affected by parameters like category (genre), sequel, star cast, number of screens, sentiment expressed in tweets and average tweets received (hype). We then propose a regression model that can be used to forecast the box office revenue earned by a movie on its opening weekend as well as can be used to predict the total box office gross of a movie.

**Keywords** Film, Box Office Prediction, Twitter, Regression, Predictive Model

**1 Introduction**

According to a study by the U.S. Bureau of Economic Analysis and National Endowment, creative industries led by Hollywood account for about USD 504 billion, or at least 3.2 per cent of US goods and services (Hollywood Reporter, 2013). By comparison, the arts and culture sector outpaced the U.S. travel and tourism industry, which was 2.8 percent of GDP in 2011, based on the federal estimate. This finding surprises us. It therefore becomes important to study the effects of a movie release on our society.

Unpredictability of the movie demand makes the movie business one of the riskiest endeavors for investors to take in today’s competitive world. Big budget films are at the risk of making revenue, they need to have an estimate of how much the movie will make and if it is less than expectation what measures the crew should take in order to increase the hype. Forecasting has always been a difficult and challenging problem that’s because movies generate income from several revenue streams including theatrical exhibition, home video, television broadcast rights and merchandising. However, theatrical box office earnings are the primary metric for trade publications in assessing the success of a film, mostly due to the availability of the data compared to sales figures for home video and broadcast rights, and also due to historical practice (List of Highest Grossing Films, Wikipedia). The opening weekend of a movie’s release typically accounts for 25% of the total domestic box office gross, so we would expect that the opening weekend’s grosses would be highly predictive for total gross (Jeffrey S. Simonoff, 2000).

Our goals in this paper are as follows. First, we consider well known parameters and try to prove or disprove the common wisdom. Aspects of movies like category, star cast, sequel, production budget are considered. But these aspects don’t contain opinions of the mob therefore may produce less accurate results. To increase accuracy we include opinions of users along with aspects of movie. Our mindsets are now influenced by what we gather through social networks. Social media plays a vital role in influencing mindsets and so we decide to investigate in its power to improve predictions. The topic of movies and entertainment in general, is of considerable interest among the social media user community. A large number of Twitter users are discussing on movies and have a substantial variance in their opinions (S. Asur, 2010). Therefore, we consider using social media, particularly Twitter1, to further explore correlations. Finally, we propose two regression models that can be used for predicting the Box Office revenue of movies.

**2 Related Work**

Despite the difficulties associated with the unpredictable nature of box office revenue, several researches have attempted to develop forecasting models. J. Simonoff and I. Sparrow (2000) have studied parameters affecting box office revenue of movies, but however they did not get sufficiently accurate results (*R2=44%*). A. Kennedy (2008) has tried to predict revenue of movies using reviews on blogs and critical reviews by practical critics but not using social media. The research focuses on using movie reviews as a parameter in developing forecasting models. But a prominent limitation is that there can be different sources of reviews and not consistent across a distribution. S. Asur and B. Huberman (2010) have used Twitter to propose a generic model for forecasting. They have shown that how social media contains insights which can be used to develop a forecast model. They focus completely on social media and have failed to consider movie aspects.

M. HUANG[6] has used Word-Of-Mouth model to forecast box office for movies. But this paper too doesn’t seem to use Social Media to forecast. **Other attempts by Ramesh Sharda (2006), Nikhil Apte, Mats Forssell, Anahita Sidhwa [8] also lack modern social media research complementing movie aspects.**

**3 Data Source**

Information from over 100 movies has been used to explore box office revenue and to verify the accuracy of our research. Movie information like release date, genre, opening weekend revenue, opening weekend theatre count, total domestic gross, widest release theatre count, movie summary, production budget etc were obtained from trusted sources including BoxOfficeMojo2, IMDB3, the-numbers4 and have been verified for consistency.

We used the Twitter Streaming API5 to gather tweets. We connected our server 24\*7 to the Twitter Streaming end-point for a period. To ensure that we obtained all tweets referring to a movie, we used keywords present in the movie title along popular hashtags for the movie as search arguments. Tweets arrived at our server as they were posted by users and were stored in a MySQL database ensuring that we had the timestamp, user screen name, tweet text for our analysis. We gathered over 2 million tweets of all languages referring to 20 different movies released over a period of 2 months.

We ensured data cleaning through spelling and grammar correction, and other sanity checks using GATE tool. Twitter by-default filters spam tweets, but we also drop tweets from users whose follower count is lesser than our threshold value. For knowing the follower count we query to the user’s profile using the Twitter Search API6. Also tweets pertaining to two or more movies, repeated tweets from same user are dropped.

We calculated polarity of tweets using Stanford NLP sentiment analysis tool7. The polarity was assigned in range of 0 to 4, with range as 0 for most negative and 4 for most positive. The tool was trained on movie reviews particularly. Stanford NLP tools works on sentence level but there may be two or more sentences in a tweet, in that case we used the sentiment expressed by the longest sentence.

**4 Exploration of Parameters**

Before proposing a forecasting model, we first explore the parameters that are correlated to Box Office Revenue of movies.

4.1 Category

We now verify the common wisdom that the category of movie is a parameter affecting box office revenue. For our analysis purposes we consider over 100 Hollywood movies.

We apply single factor Analysis of Variance (ANOVA) test to determine if there is difference between means of opening weekend box office revenue of categories for statistical significance. We have considered major 10 categories for analysis.

*A. Opening Weekend Revenue*

Null Hypothesis H0: There is no difference in means of categories

H0: µ1 = µ2 = µ3 = µ4 = µ5 = µ6 = µ7 = µ8 = µ9 = µ10

H1: At least one of them is different

The summary of the data used in ANOVA test is shown in Figure 1.

The ANOVA test results are shown in Figure 2. From Figure 2, we see that the F value is greater than F critical value. This signifies that our hypothesis is true and there is a difference in means of groups. We can safely conclude that category can be one of factors affecting opening weekend box office revenue.



Figure . Summary of opening weekend Data for ANOVA test



Figure ANOVA Test results

*B. Total Domestic Box Office Gross*

On similar lines, we perform ANOVA test to reject our null hypothesis that means of several categories are equal. From figure 3 we see that F value if greater than F critical value. This proves our hypothesis that means of total box office gross of different categories of movies are different.



Figure Summary of total gross data for ANOVA test



Figure ANOVA test results

Thus, from both our results, we conclude that category is one of parameter correlated to box office revenue of Hollywood movies.

4.2 Sequel

It is a common belief that movies that are sequels, remake or reboot of previous movies produce higher box office revenue. So we decided to investigate in it. We consider two categories of movies; first category consists of movies that are sequel, remake or reboot, and the second category consists of movies that are first installments of their series or others. We perform z-tests to see that if mean of sequel movies are higher than mean of non sequel movies.

*A. Opening Weekend Revenue*

We now perform t-test on our samples to prove our hypothesis. The first category consists of opening weekend revenue of sequel movies and second consists of opening weekend revenue of other movies that are not sequel, reboot or remake.

Null Hypothesis H0: Means of opening weekend revenue of the two categories are equal

H1: Means of the categories are different

The results of t-test are as given in figure 5



Figure z-test results

Here we see that t Stat < t critical two tail, that’s why we cannot reject the null hypothesis. The observed difference between the sample means is not convincing enough to say that the average opening weekend revenue of sequel movies and not sequel movies differ significantly.

*B. Total Domestic Box Office Gross*

We again perform the same test but this time considering the total box office gross of movies instead of opening weekend revenue.



Figure t-test results

Here we see that t Stat < t critical two tail, that’s why we cannot reject the null hypothesis. The observed difference between the sample means is not convincing enough to say that the average opening weekend revenue of sequel movies and not sequel movies differ significantly.

Therefore, the common belief that sequels generate higher revenues is contradicted. Therefore use of sequel as a parameter in forecasting revenue is not justified. In other terms, sequel is not a major factor affecting box office revenue.

*4.5Production Budget*

We now try to find if production budget of a movie is one of factor that can be considered in our predictive model. The correlation between opening weekend revenue and production budget is shown in figure. The coefficient of correlation is 0.68. The correlation between opening weekend revenue and production budget is shown in figure. The coefficient of correlation is 0.65

Though we see a positive correlation in both cases there are certain difficulties which prevents us from using it in our predictive model. First, most of budget figures are difficult to obtain. The figures available are not complete and are usually cloaked in secrecy. For example, budgets may ignore promotion, distribution and advertising costs. Consequently, it is impossible to exactly know the production budget. For the above mentioned reasons we won’t include it in our predictive model.

*4.3 Star Value*

A superstar is defined as one who contributes to the up-front in the movie. With some actors receiving as much as USD 25 million per movie, producers know that a cautious choice to selecting lead actors will affect their profit largely. Ravid (1999) has found conflicting results with respect to contribution of superstars to the revenue earned by the movie.

We now test whether a stars’ Twitter follower count is correlated to the box office revenue of movies he or she is in. Almost all stars have a Twitter account but about 80% of them don’t have a verified account8. Due to unverified accounts, and existence of multiple accounts of the same name, no consistent scale can be setup. Thus, on analysis we found that majority of cast in the movie including off-screen, do not have a verified Twitter account. For example, as of Feb 2015, to our surprise, the superstar Bradley Copper does not have a verified Twitter account. Due to unavailability of consistent data, no correlation between Twitter follower count and box office revenue can be established.

Thus we rely on traditional method of analyzing value of stars. Anita Elberse (2007) has proposed that a movie’s box office revenue depends on the star’s economic reputation as well as his or her reputation, reflected by his or her awards or nominations. We do see the significance of the director due to the fact that the most earlier studies that included star value of the movie director did not find it to be a significant contributor (Litman, 1983; Litman & Kohl, 1989; Sochay, 1994).

*4.4 Number of Screens*

Number

Our research has shown close correlation between a movies’ financial success and the number of screens in which the movie was shown during its initial launch. The results are shown in table.



Table 1 coefficient of correlation between parameters

OW: Opening Weekend

THCNT: Opening Weekend Theatre Count

WIDEST THCNT: Widest Theatre Count



Table 2 coefficient of determination between parameters

*4.6 Rate of Tweets*

Sitaram Asur and Bernardo Huberman have shown that the social media can be used to make market based prediction. Making the hypothesis “A movie well talked about, is well watched”, we independently carried out research to see that if the rate of tweets per hour affect the box office revenue of movies.

For opening weekend using the average tweet rate, the results are shown in table below.

|  |  |  |
| --- | --- | --- |
|  | R2 | p-value |
| Opening weekend revenue | 0.80 | 3.78e-09 |
| Total gross | 0.59\*\*\*\* | 4.32e-04\*\*\*\* |

**4 Forecasting Model**

We now propose forecasting models using multiple linear regression method based on the parameters analyzed.

*A. Opening Weekend Prediction*

Tweets of movies start pouring in quite before the release of movie. Prior to release of movie, media companies and producers generate promotional information in the form of trailer videos, news, blogs and photos. We are interested in finding the relationship between the revenue of movie and factors like tweets per hour, distribution factor sequel etc.

The model proposed by Sitaram Asur and Bernardo Huberman is extended here to forecast box office revenue of Indian movies on opening weekend.

The proposed regression model is as follows:

I = βa\*A + βd\*D + βc\*C + βe\*E + ε

Here,

I: Opening Weekend Income

A: Avg. tweets per hour

D: Distribution factor i.e. number of theatres movie is released in

C: Category of movie (Dummy variable)

E: Star cast (Rating: A+, A, B, C)

ε: Error factor

*B. Total Gross*

After the release of a movie, the sentiments of people come into play. We are influenced by what our network talks about. Due to difference in sentiments expressed by users, we decide to include the sentiment expressed in tweets in our model. To calculate the sentiment of tweets we used Stanford NLP tool. To forecast total gross, an additional polarity parameter is added to the model.

The proposed regression equation is as follows:

G = βa\*A + βp\*P + βd\*D + βc\*C + βe\*E + ε

G: Total Gross

A: Avg. tweets per hour

P: Average polarity calculated (Range 0 to 4)

D: Distribution factor i.e. number of theatres movie is released in

C: Category of movie (Dummy variable)

E: Star cast follower count (Rating: A+, A, B, C)

ε: Error factor

Here, β are regression coefficients.

**5 Conclusion**

In this study, we see that different categories of movies have difference in average opening weekend revenue as well as in total box office gross. Thus, category is one factor correlated to movie revenue. Next, our results contradict popular belief that sequel of movies produce on average greater revenue. Thus sequel cannot be used as a parameter in any forecasting model. The result that star participation positively affects movies’ revenues is in line with conventional wisdom. Using the chatter from Twitter.com, we constructed a multiple linear regression model for predicting opening weekend box office revenue of movies in advance of their release. We concluded that from Twitter only average Tweets received and sentiment expressed in tweets can be used in forecasting. We further proposed a multiple linear regression model to forecast total box office gross can accurately predict box office revenue.

**Notes**

1 <https://www.twitter.com>

2 <https://dev.twitter.com/streaming/overview>

3 <http://www.boxofficemojo.com>

4 <http://imdb.com>

5 <http://the-numbers.com>

6 <https://dev.twitter.com/rest/public/search>

7 <http://nlp.stanford.edu/sentiment>

8 [https://support.twitter.com/articles/119135-faqs-about-verified-accounts#](https://support.twitter.com/articles/119135-faqs-about-verified-accounts)

**References**

[1] The Associated Press, Hollywood Reporter, <http://www.hollywoodreporter.com/news/hollywood-creative-industries-add-504-662691> Accessed 01 Feb 2015.

[2] List of highest Grossing Films, Wikipedia, <http://en.wikipedia.org/wiki/List_of_highest-grossing_films> Accessed 01 Feb 2015.

[3] Predicting movie grosses: Winners and losers, blockbusters and sleepers, Jeffrey S. Simonoff and Ilana R. Sparrow, Chance, 13(3), 15-24 (Summer 2000)

[4] Predicting Box Office Success: Do Critical Reviews Really Matter?, Alec Kennedy, University of California, Berkeley, **##2008##**

[5] Predicting the Future with Social Media, Sitaram Asur, Bernardo A. Huberman, HP Labs, CA

[6] Using Online WOM to Forecast Box Office for Movies Coming Soon, Minxue HUANG, Feng WANG, Wuahn University, China

[7] Predicting box-office success of motion pictures with neural networks, Ramesh Sharda, Dursun Delen, ElseVier

[8] Predicting Movie Revenue, Nikhil Apte, Mats Forssell, Anahita Sidhwa

[9] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang,

Christopher D. Manning, Andrew Y. Ng and Christopher Potts, Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank

[10] Predicting box-office success of motion pictures with neural networks, Ramesh Sharda, Dursun Delen

[11] Elberse, Anita (2007), “The Power of Stars: Do Star Actors Drive the Success of Movies?” Journal of Marketing, 71(4), 102–120.

[12] Litman, B. R. (1983). Predicting success of theatrical movies: An empirical study. Journal of Popular Culture, 16(9), 159–175.

[13] Litman, B. R., & Ahn, H. (1998). Predicting financial success of motion pictures. In B. R. Litman (Ed.), The motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc.

[14] Sochay, S. (1994). Predicting the performance of motion pictures. The Journal of Media Economics, 7(4), 1–20.