Movie Box-Office Gross Revenue Estimation

Shaiwal Sachdev, Abhishek Agrawal, Shubham Bhendarkar, Bakshi Rohit Prasad and Sonali Agarwal

Abstract In this research work, movie box-office gross revenue estimation has been performed using machine learning techniques to effectively estimate the amount of gross revenue a movie will be able to collect using the public information available after its first weekend of release. Here, first weekend refers to first three days of release namely Friday, Saturday, and Sunday. This research work has been done only for the movies released in USA. It was assumed that gross revenue is equal to the amount of money that is collected by the sale of movie tickets. Data collected has been collected from IMDB and Rotten Tomatoes for movies released from the year 2000–2015 only. Multiple linear regression and genre-based analysis was used to effectively estimate the gross revenue. Finally, Local regression methods namely local linear regression, and local decision tree regression were used to get a better estimate.

Keywords Gross revenue · Box-office · Linear regression · Decision tree regression · IMDB · Rotten tomatoes · Machine learning

S. Sachdev (✉) · A. Agrawal · S. Bhendarkar · B. R. Prasad · S. Agarwal
e-mail: sonali@iiita.ac.in
Indian Institute of Information Technology Allahabad, Allahabad, India
e-mail: iit2013196@iiita.ac.in

A. Agrawal
e-mail: iit2013128@iiita.ac.in

S. Bhendarkar
e-mail: iit2013172@iiita.ac.in

B. R. Prasad
e-mail: rs151@iiita.ac.in

S. Agarwal
e-mail: sonali@iiita.ac.in

P. K. Sa et al. (eds.), *Recent Findings in Intelligent Computing Techniques*,
Advances in Intelligent Systems and Computing 709,
https://doi.org/10.1007/978-981-10-8633-5_2

1 Introduction

Film industry is a big business in United States. It is one of the biggest players in the entertainment industry. Predicting the gross revenue of a movie has become need of the hour. Lot of researchers have used different models but still there is no computational model that can effectively predict the box-office gross revenue movie will be able to collect. This depends on a lot of factors like number of available theater screens, budget of the film, star cast, genre of the movie, MPAA rating (Motion Picture Association of America film rating system), and release year. This paper uses a combination of regression methods and specific genre-based method using the revenue data collected from IMDB and Rotten Tomatoes of past 15 years movies to estimate the gross revenue after its first weekend of release. This method can be used by movie producers and production studios who by looking at estimated values of revenue can take different steps on deciding the budget for things like marketing and promotion. Also can be used by movie theaters as they can also estimate their sale of tickets. If the estimated revenue is lower than expected, studios may increase their promotion budget and may even think of releasing the movie outside the domestic space. In these cases, studios may try to release it in more theaters and invest more in advertisement. The goal of this research paper is to propose a method that will be able to effectively estimate the box-office gross revenue for a movie using the public information available after its first weekend of release.

2 Related Work

A lot of factors that affect the revenue prediction model have been studied and used by different researchers in combination with multiple machine learning algorithms. Do Critics Reviews Really Matter? As mentioned in [1], Eliashberg and Shugan analyzed the impact of critic reviews on box-office success. They divided the role of critic into two dimensions, influencer, and predictor. Critic is said to be an influencer, when his or her reviews influence the box-office results. And to be a predictor when he or she is able to predict the success of movie based on reviews. Ravid in [2] concluded that box-office revenue increases with increase in positive reviews by reviewers. However, according to Reinstein and Snyder in [3], few critics have the power to manipulate the box-office revenue hence, only critic ratings cannot judge the financial success. Research work of Litman [4] used multiple linear regression to predict the box-office success. He concluded that star cast, genre of movie, MPAA rating, and release date are all determinants of the financial success of the film. Forswell in their work [5] used linear regression model on features like Opening Weekend revenue, budget of the movie and number of theater screens. Robert in [6] divided the feature set into three types; simple which is numeric only, complex that is numeric, and text and sentiment which includes all possible types. They used Logistic regression for classification. Simonoff in [7] used different parameters like

Fig. 1 Proposed methodology

production budget, whether movie is a sequel, star power, etc., that can predict the box-office revenue before release. Marton and Taha in [8] tried to do early prediction of movie success using the Wikipedia activity Big Data. Big Data analytics and computing is one of the current focus of research these days in multiple domains [9, 10]. Vitelli in [11] tried to create a set of features and did extract values from graphical properties of the actor–actor graph, actor–movie graph, and movie–movie graph relationships. Dursun Delen in [12] used neural network with features like MPAA rating, genre of movie, star value, and sequel importance for prediction of prerelease revenue. Work by Anast [13] related genre to movie attendance and concluded that violence and eroticism have a positive correlation with movie attendance. Unlike previous works Ryan as in [14] tried to estimate the foreign box-office revenue which depends on domestic success, language adaptability, cultural differences, MPAA rating, etc. Thorsten in [15] did a study whether the success of one feature can affect the effect of other on revenue or not.

Considering related research, we find that none of them did the genre-based analysis. Moreover, all features cannot be a determinant for all types of movies. So, this paper applies a genre-specific approach along with regression models to effectively estimate the box-office gross revenue. The proposed methodology is shown in Fig. 1.

3 Proposed Methodology

3.1 Data Collection

Movie related data was collected from the two most popular websites IMDB and Rotten Tomatoes for the past 15 years (year 2000–2015). Movies collected consisted of many genres namely Action, Adventure, Animation, Thriller, Horror, Sci-Fi, etc.

Field Information. Within the collected data, some of the general features were the title of movie, genre of movie, release date, budget of the movie(USD), number of screens it was released in the opening weekend, opening weekend revenue(USD), IMDB user rating, TomatoMeter, TomatoRating, UserMeter, UserRating, Popularity from IMDB, Rotten Tomatoes, and gross domestic revenue for the movie in USA.

Field Description. TomatoMeter is based on the published opinions of hundreds of film critics and is a trusted measurement of movie quality for millions of movie-

goers. Meter represents the percentage of critic reviews that were rated above 5 out of 10. TomatoRating is the average critic rating on a ten point scale.

UserMeter is percentage of users who rated it positive. UserRating is the average user rating on a five point scale. IMDB user rating is done by users or viewers on a ten point scale. Popularity score from Rotten Tomatoes gives us the number of users who wanted to watch the movie and who watched and rated it. Popularity count from IMDB is the number of movie page views for that week.

3.2 Data Cleaning

Movies with insufficient information were removed from the dataset. Budget of the movies was in different currencies. Currency Conversions was done to convert all budget values into USD.

3.3 Data Transformation

Box-Office gross revenue is simply the total sum of the money collected by the sale of movie tickets. Average movie ticket prices have increased rapidly each year rising from 5.39 USD in 2000 to around 8.66 USD in 2016. To adjust the revenue data collected namely budget of the movie, opening weekend revenue, and box-office gross revenue, the change in ticket price was considered as an inflation parameter and all the revenue was changed according to ticket prices in 2016.

Standardization. Various features in the dataset are brought together to the same level or scale. This will make all the different features having different ranges to contribute proportionally to the final gross revenue. In fact, the ranges are unbounded, budget of the movie or box-office gross revenue has no maximum or minimum value. After performing standardization, all the features will have zero mean and unit variance. Each sample and feature is standardized using Eq. 1.

$$x_1 = \frac{x - mean}{Standarddeviation} \tag{1}$$

4 Estimation Using Regression-Based Modeling

Mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD) is the mean of the percentage error for each sample. Let A denote the actual value and F denote the predicted value. Let n number of test movies, then Eq. 2 specifies the calculation of MAPE.

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{2}$$

Assuming linear relationship between gross revenue and all the other features, multiple linear regression was used. Using this model and all the features, error (MAPE) was around 110%. Equation 3 describes the linear model where W is the weight and b is the bias.

$$Y = W1 * Budget + W2 * OpenWeek + W3 * Screens + W4 * userRating + \cdots + b \tag{3}$$

4.1 Training Split on the Basis of Genre

Each genre of movie shares a different relationship with each of the features. For example, budget of the movie maybe important for Action or Animation movies but not for a Horror movie. UserRatings or UserReviews has more role to play in the prediction model for movies which are of documentary or Horror type and lesser on Adventure or Comedy movies. Whole training set is split on the basis of genre of movie. Then feature selection algorithms are used to find the best set of features for each genre. This way, model will adapt to different genres of movies and use the features accordingly.

4.2 Feature Selection

Feature selection is used basically for two reasons:

- To avoid over fitting by reducing number of features and to improve generalization of model.
- To gain better understanding of features and their relationship to response variable.

Among various techniques best subsets regression guaranteed the best result in our work. It selects best set of features based on statistical criteria; Mean absolute percentage error (MAPE). Result obtained after best subsets regression method is given in Table 1. Here, OpenWeek refers to the opening weekend revenue and Popul refers to popularity. Result shows that the budget of the movie does not affect the gross revenue of horror, documentary, Sci-Fi, and history movies. Opening weekend is the most important feature that is present in all of the genres. Critic rating and popularity do not seem to affect the performance of a lot of movies. Only documentary, drama, history, and music were affected a bit. Multiple linear regression model is used for each genre of the movie. Here, training data is the set of movies released between the year 2000–2013 and test data is of the movies released from the year 2014–2015. Test set consists of 400 movies.

Table 1 Best combination and MAPE for each genre

Genre	Best combination	MAPE
Action	OpenWeek, Popul(IMDB)	46.45
Adventure	OpenWeek, budget	49.16
Animation	OpenWeek, budget	36.69
Drama	OpenWeek, budget, Popul(IMDB)	95.45
Comedy	OpenWeek, budget	50.26
Sci-Fi	OpenWeek	24.20
Romance	OpenWeek, budget, Popul(Rotten)	90.45
Music	OpenWeek, budget, UserMeter, Popul(Rotten)	60.70
Fantasy	OpenWeek, budget, screens, Popul(IMDB), UserRating	47.90
History	OpenWeek, Popul(IMDB), TomatoRating	62.38
Documentary	OpenWeek, Popul(IMDB), TomatoRating, UserRating	42.57
Horror	OpenWeek, Popul(Rotten)	23.47
Mystery	OpenWeek, budget, Popularity(IMDB), Popularity(Rotten)	49.89

5 Multi-genre Analysis

The movies in the test set were having more than one genre. To handle such type of data, linear model was run for each genre using the best set of features found using best subsets regression. Then final estimated box-office gross revenue is the arithmetic mean of gross revenue estimated by the model for each genre. Error (MAPE) was found to be 53.96%.

5.1 Training Split on the Basis of TomatoRating and Screens

After splitting the training set on the basis of genre, then it was split on the basis of TomatoRating and number of screens as shown in Fig. 2. Then linear regression model is run and error (MAPE) is calculated as shown in Table 2.

For TomatoRating, we observed there are two type of movies

Type1, High Critic Rating > 5.

Type2, Low Critic Rating ≤ 5.

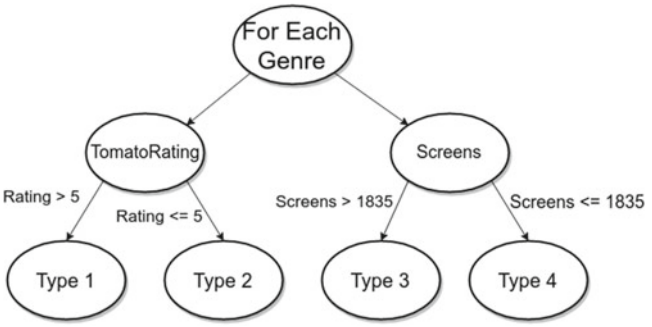


Fig. 2 Training split on the basis of TomatoRating and screens

Table 2 Linear regression result after split for test set

Split	All	High	Low
Rating	34.18 (400)	42.6 (244)	24.25 (156)
Screens	39.92 (400)	26.96 (210)	127.32 (190)

For Number Of Screens, we observed there are two type of movies:
Mean value of number of screens taken from the Training Set is 1835.
Type3, High Critic Rating >1835.
Type4, Low Critic Rating $\leq$ 1835.

5.2 Local Regression Models

Neighbor Search. For each test movie, nearest samples in the training set are found out. Two tuples are made on the basis of genre of movie like for Adventure movies, (Opening Weekend, Budget) will be the tuple. Now, Euclidean distance between the test tuple and each of the training data tuple is calculated and 50 nearest ones are considered to be the training set. Multiple linear regression and decision tree regression were run using these 50 neighbors as the training set. In Table 3, movies with higher number of screens (that is less than 1835) gave better result whereas in Table 4, movies with lower rating gave better result. But, on comparing the two tables, movies with higher rating performed better than movies with lower number of Screens.

On observing the two tables, model performs poorly on movies with low number of screens (that is less than 1835). To overcome this, two-way split is used. When the movie was released on screens > 1835, model should use the Type3 training set. When the movie is released on screens $\leq$ 1835, model should use split on the basis of TomatoRating and then use Type1 or Type2 training Set accordingly. This approach is shown in Fig. 3.

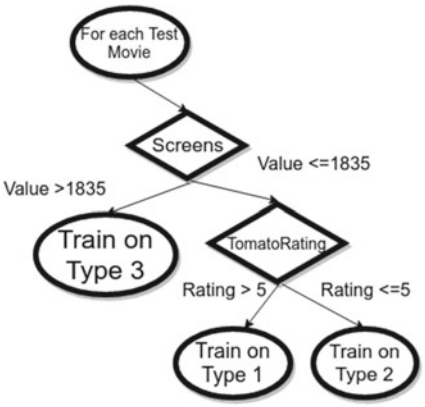
Table 3 Local models (on the basis of splitting by screens)

Algorithm	All (400)	High (210)	Low (190)
Linear regression	37.966	18.6	84.9
Decision tree regression	25.77	11.8	50.6

Table 4 Local models (on the basis of splitting by TomatoRating)

Algorithm	All (400)	High (244)	Low (156)
Linear regression	32.47	41.38	23.21
Decision tree regression	28.78	29.17	17.04

Fig. 3 Final approach



Using the approach in Fig. 3, local decision tree regression gave the best result. Error (MAPE) for the test set of 400 movies was found to be 24.76%.

6 Conclusion

Presented work performs genre-based splitting of training set and further division on the basis of number of screens. If the number of screens is less than mean value 1835, then we use Type 1 or Type 2 training set. Error is minimum when movies with higher number of screens (210 movies) were tested. This was as low as 11.8% Using this combined type of approach (Fig. 3) and Local decision tree regression algorithm, error (MAPE) for 400 test movies was 24.76%.

Table 5 Test on recent 5 releases of 2016

Name	Real gross	Predicted gross	MAPE (Percent)
Conjuring 2	102,461, 593 USD	117,240,783 USD	14.42
The Angry Birds	107,506,776 USD	145,956,013 USD	35.5
Deadpool	363,024,263 USD	358,836,741 USD	1.15
The Legend of Tarzan	126,585,313 USD	113,305,012 USD	10.49
The Jungle Book	363,995,937 USD	425,317,712 USD	16.8

7 Test on Recent Releases of 2016

Some recent releases of 2016 were tested and results are shown in Table 5.

References

1. Eliashberg, J., Shugan, S.M.: Film critics: influencers or predictors? J. Mark. 68–78 (1997)
2. Ravid, S.A.: Information, blockbusters, and stars: a study of the film industry. J. Bus. **72**(4), 463–492 (1999)
3. Reinstein, D.A., Snyder, C.M.: The influence of expert reviews on consumer demand for experience goods: a case study of movie critics. Working Paper, University of California-Berkeley and George Washington University (2000)
4. Litman, B.R.: Predicting success of theatrical movies: an empirical study. J. Pop. Cult. **16**(4), 159–175 (1983)
5. Apte, N., Forssell, M., Sidhwa, A.: Predicting Movie Revenue. CS229, Stanford University (2011)
6. Yoo, S., Kanter, R., Cummings, D.: Predicting Movie Revenue from IMDb Data. Stanford University (2011)
7. Simonoff, J.S., Sparrow, I.R.: Predicting movie grosses: winners and losers, blockbusters and sleepers. Chance **13**(3), 15–24 (2000)
8. Mestyn, M., Yasseri, T., Kertsz, J.: Early prediction of movie box office success based on Wikipedia activity big data. PloS One **8**(8), e71226 (2013)
9. Prasad, B.R., Agarwal, S.: Comparative study of big data computing and storage tools: a review. Int. J. Database Theory Appl. **9**(1), 45–66 (2016)
10. Prasad, B.R., Agarwal, S.: Stream data mining: platforms, algorithms, performance evaluators and research trends. Int. J. Database Theory Appl. **9**(9), 201–218 (2016)
11. Predicting Box Office Revenue for Movies: Matt Vitelli (2015)
12. Sharda, R., Delen, D.: Predicting box-office success of motion pictures with neural networks. Expert Syst. Appl. **30**(2), 243–254 (2006)
13. Anast, P.: Differential movie appeals as correlates of attendance. J. Q. **44**(1), 86–90 (1967)
14. de Silva, B., Compton, R.: Prediction of foreign box office revenues based on wikipedia page activity (2014). arXiv:1405.5924
15. Hennig-Thurau, T., Houston, M.B., Walsh, G.: Determinants of motion picture box office and profitability: an interrelationship approach. Rev. Manag. Sci. **1**(1), 65–92 (2007)
16. Internet Movie DataBase: http://www.imdb.com/
17. Rotten Tomatoes: https://www.rottentomatoes.com/