Predicting Bollywood Movies Success Using Machine Learning Technique

Garima Verma[1], Hemraj Verma[2]

[1,2]Department of Information Technology, Head, Faculty of Management
[1]Garimaverma.research@gmail.com, [2]Hemraj77@gmail.com
DIT University, Dehradun, INDIA

Abstract: The main purpose of this paper is to develop a model for predicting the success of movie (Bollywood movie) being a Hit or Flop, long before a movie is actually released. Using Logistics Regression algorithm, a supervised machine leaning algorithm, a model has been developed for predicting Hit Bollywood movies. The model developed has been able to predict the Hit movies with an accuracy up to 80%. The proposed model is unique in the sense that it uses MusicRating of a movie as a predictor which is a unique feature of Bollywood movies. The study holds lot of relevance to practitioners as well as to academicians as the former may benefit by using this model to predict the success of a Bollywood movie before its release and latter may gain new insights into this body of knowledge.

Keywords: Prediction models, Logistics regression, Bollywood movie, Predictors

I. INTRODUCTION

Bollywood, is the India's biggest movie industry based in Mumbai. It produces maximum number of Hindi movies per year on an average of 250 to 300 movies per year [1]. This figure is really higher than any other country's movie industry. Bollywood is a multi- billion dollar industry, which is growing day by day approximate 11% annually [2]. But there are very less movies which get success and win hearts of audience, rest generally come in the category of flops or average [2]. This gives us an opportunity to analyze the movie data and do the prediction about the success of movie before a movie's release. This can be done on the basis of various parameters such as cast of the movie, director of the movie, release date of the movie, genre, quality of script, Movie ratings etc. Although various studies have been done for the Hollywood movies but the parameters taken in these kind of studies are not applicable directly to Bollywood movies, largely due to the type of interest and contents of the movies. For example, Bollywood movies have songs which becomes very important parameter for the entertainment for audience in India. In fact it is presumed that it can be a one of the parameters in the success of the movie, but Hollywood does not have this parameter [3][4].

The main objective of this paper is to propose a model that can predict the success or failure of a Bollywood movie before its release. The study will benefit all the stakeholders who are involved in the movie production, right from proposal of an idea till its release. Further, this model can be helpful to all those who are new to this industry and wish to take all necessary precautions for making a successful movie.

II. LITERATURE REVIEW

A lot of research has been done in the past to predict movie success, especially at global level, involving different types of data sources, data analysis software and algorithms for model development. In [4] authors have collected data from social sites specially twitter and face book to collect the movie data. The authors have used sentimental analysis to analyze the performance of the Indian movies at box office. In [1] authors collected data from mainly cinemalytics and BoxofficeIndia and youtube. The authors did an analysis on a software called Weka [5], which is a specially build software for applying machine learning techniques. They have used bagging technique to extract the best results. In [6], authors have done analysis for Hollywood movies success and whether it can be referred for academy awards like Oscar or not. Authors collected data mainly using Internet Movie Database (IMDb) and Boxoffice. The model predicted nine Oscar nominations correctly just before it actually happend. In [7], authors has developed a mathematical model. They have taken some parameters such as – budget, actor, director, locations, story, and release date etc. for success and failure analysis. In [8] authors have performed analysis of online movie resource of over 390,000 movies and television shows on the basic of various criteria's. In 2009 some researchers worked on movie prediction using new analysis [9] and they determined in the study that the news data analysis is as good as the analysis done on any data extracted from any other data source. Even they claimed that they have achieved better performance. In 2016, Michael has done a study and proposed a decision support system. The proposed system predicts the success or failure of a movie based on profit earned by taking the benefit of past data of 11 years from various sources [10].

III. METHODOLOGY

The main purpose of this study was to develop a model that could predict the success (called as Hit) or failure (termed as Flop) of a movie before the actual release of a movie.

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

There are several techniques that can be used to develop prediction models such as Multiple Regression, Discriminant Analysis, Artificial Neural Network, Logistics regression, Multinomial regression etc. However, Logistics regression (LR) is one of the most popular algorithm that can be used to predict a binary outcome and one or multiple continuous or categorical predictor variables. In fact, LR can be used for outcome variable with more two categories as well [11]. Since LR uses logit model therefore the value of outcome variable ranges only between 0 and 1. More importantly, LR does not follow the assumption of normality of sample data thus making it much more flexible to use it for prediction. [12]. IBM SPSS21.0 has been used to process and analyze the data and develop LR model.

A. Feature Selection and Pruning

Though there are several models that have used various features for movie prediction such as Movie Budget, Lead Actor Rank, Director Rank, No. of Screens used for release of the movie, Movie Ratings, Music Ratings etc. However, due to complexity in data collection and keeping the model simple the proposed LR for Bollywood movie prediction used three predictors such as *Total number of screens used for a movie, IMDb rating (Internet Movie database rating of Bollywood movies)* and *Music Rating* of a movies (A feature unique to Bollywood movies). The model has included *movie verdict* an outcome variable to predict movie success (Hit =1) or failure(Flop-0).

B. Data Collection:

The data for the study has been collected from variety of online sources such as imdb.com, bollymoviereviewz.com, planetbollywood.com, boxofficeindia.com, and bollywoodhungama.com [13][14][15][16][17]. The data was largely collected through web scrapping using scraper extension in Chrome. However, not all the data could be fetched through scraper. Therefore, some of the data was collected looking for individual movie information spread over many websites. Especially, data related to music rating of movies was mostly collected in bits and pieces from different websites.

C. Data Pruning:

Initially a list of more than 2000 movies was extracted through scraper but it contained information related to only movie ratings and outcome variable i.e. Hit or Flop movie. Therefore, the data related to second feature i.e. number of screens for release of a movie generated a second list of around 400 movies and was combined with movie rating. However, the third feature related to music rating of a movie could only generate 116 movies. Therefore, All missing data points were excluded and our final list of movies contained data related to 116 movies for one outcome variable (Hit or Flop movie) and three predictors (No. of Screens, IMDb rating, Music Rating).

IV. DATA ANALYSIS

The table-1 describes the total number of cases with number of flops (code as 0) and hits (coded as 1). Overall data for 116 movies (39.7% flop movies and 60.3% hit movies) was captured and was further divided into 80:20 ratio for training and test purpose.

TABLE I: FREQUENCY OF THE CRITERION VARIABLE VERDICTCODED

		Frequency	Percent
Valid	Flop (0)	46	39.7
	Hit (1)	70	60.3
	Total	116	100

The table-2 describes the basic statistics related to all the predictor variables used for LR model development. Since there are no missing values so whole of the data is used for further analysis.

TABLE II: DESCRIPTIVE STATISTICS OF PREDICTORS

	No. of Screens	IMDb_ratings	Music Rating
Valid cases	116	116	116
Missing cases	0	0	0
Mean	1792.46	6.29	7.16
Std. Deviation	1020.75	1.75	1.19
Skewness	.813	-.722	.075
Kurtosis	-.078	-.719	-.476
Minimum	300.00	1.70	5.00
Maximum	4400.00	8.60	10.00

The table-3 tests whether the model is significant if predictors are used in LR model. The result show that the model is significant.

TABLE III: OMNIBUS TESTS OF MODEL COEFFICIENTS

		Chi-square	df	Sig.
Step 1	Step	49.933	3	.000
	Block	49.933	3	.000
	Model	49.933	3	.000

The table-4 indicates the percentage of variation explained by three predictors. The value of Cox and Snell R Square is 40.6% and Nagelkerke R Square is 54.9%.

TABLE IV: MODEL SUMMARY FOR TRAINING DATA

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	78.954^{a}	.406	.549

Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

The table-5 indicates the goodness of fit of logistic regression for the data. The Chi-square value is insignificant, indicating that LR fits the data well.

TABLE V: HOSMER AND LEMESHOW TEST

Step	Chi-square	df	Sig.
1	6.761	8	.563

The table-6 exhibits the LR coefficients with their relative significance and odds ratio. All the three predictors have been found significant.

TABLE VI: VARIABLES IN THE EQUATION

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Screens	.001	.000	4.585	1	.032	1.001
	IMDb_ratings	.848	.237	12.768	1	.000	2.336
	MusicRating	.891	.316	7.944	1	.005	2.437
	Constant	-12.570	2.698	21.713	1	.000	.000

a. Variable(s) entered on step 1: Screens, IMDb_ratings, MusicRating.

The table-7 displays the correctly percentage of classified cases for training as well as test data. The proposed model have been able to achieve an accuracy of 78.1% for training data and 80.0% for test data.

TABLE VII: CLASSIFICATION TABLE[C] FOR TRAINING AS WELL AS TEST

	Observed		Predicted					
			Selected Cases[a] *(Training Data)*			Unselected Cases[b] *(Test Data)*		
			MovieVerdict		Percentage Correct	MovieVerdict		Percentage Correct
			Flop	Hit		Flop	Hit	
Step 1	Movie Verdict	Flop	25	13	65.8	6	2	75.0
		Hit	8	50	86.2	2	10	83.3
	Overall Percentage				78.1			80.0

a. Selected cases TrainingNtest EQ 1

b. Unselected cases TrainingNtest NE 1

c. The cut value is .500

The Fig-1 shows the Receiver Operating Characteristics (ROC) curve, using three predictors. It indicates the area under the ROC curve (as mentioned in table-8) of MusicRating predictor is maximum followed by IMDbRating and Number of screens (Screens).

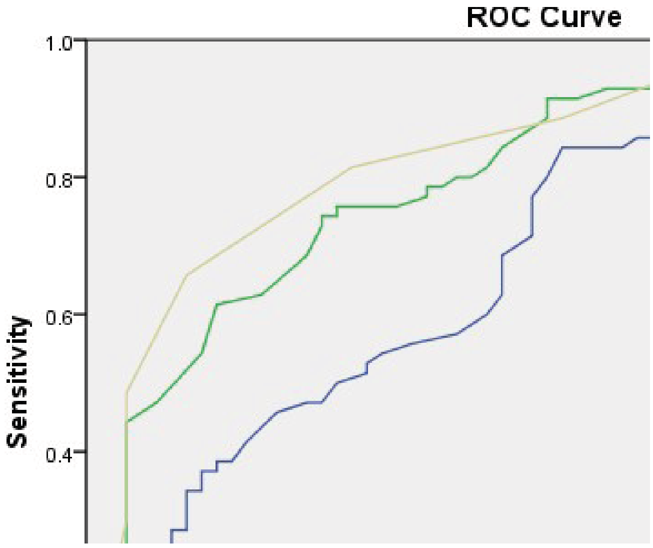

Fig. 1. ROC curve for LR model

TABLE VIII: AREA UNDER THE CURVE

Test Result Variable(s)		Area
dimension0	Screens	.623
	IMDb_ratings	.779
	MusicRating	.817

V. RESULTS AND DISCUSSIONS

The results of LR model shows that the proposed model is significant x^2 = 49.933, df =3, N= 116, p<.000 (with all predictors). This indicates that all predictors used in the model are significant in explaining the variation in outcome variable *Movie Verdict*. Further, Hosmer and Lemeshow fitness test too reveals that LR is fits the data well with x^2 = 6.761, df =8, N= 116, p =. 563. Further, table 4 exhibits that model explained 40.6% (Cox & Snell R Square) and 54.9% (Nagelkerke R squared) variation in outcome variable using three predictors. Also, the Logisttics Regression coefficient for all predictors ciz. *Screens, IMDbrating and MusicRatings* have been found significant with p values = *0.032, 0.000 and 0.005* at 5 % level of significance.

The classification table7 indicates the data analysis related to correctly classified cases for training (80% cases) and test data (20%) set. In training data, the model correctly predicted 65.8% flop and 86.2% Hit movies. Overall, 78.1% movies were accurately classified by this LR model. For Test data (Unselected cases in table-7), the classification was slightly better with 75% and 83.3% correctly classified movies as

Flops and Hits and overall accuracy being 80%. The cut-off value for classification was taken as default 0.50.

Overall the model achieved an accuracy of close to 80% which appears to reasonably good given the fact that a unique variable, music rating, was included in the model which in some way is unique to Bollywood movies. The inclusion of this variable size also created a limitation of sample size being small as not enough data is available or is spread over several sources in bits and pieces.

Table-6 indicates the logistics regression coefficients related to each of predictor with relative significance and odds-ratio. All of the three predictors have been found significant at 5% level of Alpha. MusicRating (with Beta=.891 has been found to be the most significant predictor of the movie success, followed by IMDb rating ((with Beta=.848) and number of Screens for a movie (with Beta=.001). The odds ratio for MusicRating exhibits that the probability of a movie being a hit is 2.437 times more if music of the movie is good, given other predictors are held constant. Similarly, the chances of a movie being a Hit are 2.336 more if IMDb ranking is good and 1.001 times more if Number of screens used for a movie are large, when other predictors are held constant.

Further, the above results are also exhibited by ROC curve *(Plot of 1-Specificity on X-axis and Senstivity on Y-axis)*. The area under the ROC curves of three predictors MusicRating, IMDb Rating and No.of Screens is 0.817, 0.779 and 0.623. It shows that the each of dataset related to these predictors have 81.7%, 77.9% and 62.3% concordant pairs.

VI. CONCLUSION

In this study the effort has been made to develop a simple model for predicting the success of a movie before its release using Logistics Regression algorithm. Using MovieVerdict (Hit or Flop) as outcome variable, the model used three predictors No. of screens, IMDb ratings, MusicRating of a movie as predictor variables. It was found that the three predictors chosen have significant explanatory powers to determine the success or failure of a movie with MusicRatings found to have most influence in making a movie hit or flop, followed by IMDb rating and No. of screens used for release of the movies.

ACKNOWLEDGMENT

We would like to thank Dr K. K. Raina Vice Chancellor DIT University, Dehradun for encouraging and supporting us in making it possible for writing this research paper.

REFERENCES

[1] S. R. Jaiswal, and D. Sharma, Predicting success of bollywood movies using machine mearning techniques, ACM Compute 2017, Nov, Bhopal India.

[2] D. G. Taylor and M. Levin, Predicting Mobile App Usage for Purchasing and Information-Sharing, International Journal of Retail & Distribution Management, Vol. 42, no. 8, 2014, pp. 759–774.

[3] S. White, M. Saraee, and J. Eccleston, A data mining approach to analysis and prediction of movie ratings, In The Fifth International Conference on Data Mining, Text Mining and their Business Applications. University of Salford,England, 2014, pp. 344–352.

[4] D. D. Gaikar, and B. Marakarkandy, Using twitter data to predict the performance of Bollywood movies, Industrial Management and Data Systems, Vol 115(9), 2015, pp. 1604-1621.

[5] A. Mark, H. E. Frank, and H. Ian, The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Publishers, 2016

[6] J. S. Krauss, D. Simon, K. Fischbach, P. Gloor, Predicting movie success and academy awards through sentiment and social network analysis, In the proceedings of 16th European conference on Information systems, 2008, pp. 2026-2037.

[7] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, Movie success prediction using data mining, In proceedings of 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, IIT Delhi, India

[8] M. Saraee, S. White, and J. Eccleston. ”A data mining approach to analysis and prediction of movie ratings”, Vol 33, 2004, pp. 10.

[9] W. Zhang and S. Skiena, Improving movie gross prediction through news analysis, In IEEE proceedings of International Joint Conference on Web Intelligence and Intelligent Agent Technology, Italy, 2009, pp. 301-304.

[10] T. L. Michael, and K. Zhao, Early predictions of movie success: the who, what, and when of profitability, Journal of information management system, Vol 33(3), 2016, pp. 874-903.

[11] D. G. Kleinbaum, and M. Klein, Logistic Regression, New York, Springer, 2010, pp. 1-39.

[12] M. Pohar, M. Blas, and S. Turk, Comparison of logistic regression and linear discriminant analysis, Metodoloki zvezki, Vol. 1(1), 2004, pp. 143-161.

[13] 2018. boxofficeindia. Web page. www.boxofficeindia.com

[14] 2018. Cinemalytics - The Bollywood Movie Database. Web page. https://www.cinemalytics.com

[15] 2018. IMDb Database Statistics. Web page. http://www.imdb.com/stats

[16] 2018. YouTube. Web page. https://www.youtube.com/

[17] 2018. YouTube Data API | Google Developers. Web page. https://developers.